Toward a Better Understanding of Evolving Social Networks

Ties, Triads, and Time

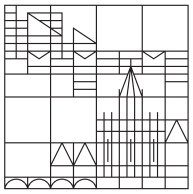
Dissertation

zur Erlangung des akademischen Grades
Doktor der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von

Bobo Nick

an der



Mathematisch-Naturwissenschaftliche Sektion
Fachbereich Informatik & Informationswissenschaft

Tag der mündlichen Prüfung:	23. Oktober 2013
Erster Referent:	Prof. Dr. Ulrik Brandes
Zweiter Referent:	PD Dr. Sven Kosub

Preface

Je pense, donc je suis

(René Descartes, Discours de la Méthode pour bien conduire sa raison et chercher la verité dans les sciences)

RETROSPECTIVELY, it was a wooden toy train system that arouse a great deal of interest in networks within a young scholar at the age of 3. In the same way it had been exciting to create interwoven train tracks in these days, long ago, I am still fascinated by the automatisms *at play* in networks today.

I thank my family for being the stable core of my ever-changing personal network through all this years, and, in particular, Julia and Caja for becoming my dearest clique therein — your love, encouragement and support made possible everything. All my friends and relatives, you are such a unique social network. Thank you!

Scientific acknowledgements. First and foremost I would like to express my gratitude to my supervisor, Ulrik Brandes, for fostering my passion for *social* networks and offering a position in his group at the University of Konstanz. In addition, there are at least seven good reasons to thank the second supervisor of my thesis, Sven Kosub, but I will black them out here...

Moreover, I would like to thank my office roommate, Jürgen Lerner, and all additional co-authors (in alphabetical order), Pádraig Cunningham, Martin Hoefer, Natalie Indlekofer and Conrad Lee (!), Martin Mader, Uwe Nagel, Steffen Rendle, Brigitte Rockstroh, and Astrid Steffen, for discussing many interesting ideas.

Dear *algos* in Konstanz, and dear *Clique* guys in Dublin, it was great fun working with all of you!

I am very grateful to Marc Scholl for joining my thesis committee. Additional thanks go to Francesca Pallotti and Britta Renner for allowing the reprint of gestalt-matrix representations of their data, Derek Greene for providing his Twitter datasets, Christoph Stadtfeld for sharing parts of his LaTeX templates, as well as Tobias Döbele and Volker Mühlberg who have been instrumental in conducting the user study on gestaltlines in the appendix.

Bobo Nick
Konstanz, 2013

Deutsche Zusammenfassung

Gegenstand dieser Arbeit sind methodische Beiträge zum besseren Verständnis sich entwickelnder sozialer Netzwerke.

Komplexe Beziehungsgeflechte zwischen sozialen Akteuren entstehen weder aus dem Nichts heraus, noch bleiben diese unverändert über die Zeit. Im Gegenteil, soziale Netzwerke sind von Grund auf dynamisch und in ständiger Entwicklung.

Das bessere Verständnis sich entwickelnder sozialer Netzwerke ist möglicherweise ein entscheidender Faktor für die Erklärung einer Vielzahl sozialer Phänomene. Tatsache ist, dass das Netzwerkparadigma immer häufiger adaptiert wird.

Die Begrifflichkeit geht zurück auf Barnes [1954]. Dort wird das Bild eines vernetzten sozialen Lebens anhand von Punkten (Akteure) und Linien (Beziehungen) umschrieben. Diese dezidiert strukturelle Sichtweise hat sich als extrem überzeugend erwiesen und spätestens mit dem Erscheinen von Google, Facebook, Twitter und Co auch außerhalb der Wissenschaft rasant verbreitet.

Obwohl Netzwerkstruktur gemeinhin durch das mathematische Konzept eines Graphen, also als die Summe einzelner Kanten, beschrieben wird, so sind soziale Netzwerke doch keinesfalls bloße Ansammlungen dyadischer Beziehungen. Vielmehr sind die beobachteten Verknüpfungen systematisch strukturiert und jenseits der dyadischen Ebene in das gesamte soziale System eingebettet. Die Natur eines sozialen Netzwerkes führt daher zu wechselseitigen Abhängigkeiten zwischen Akteuren und deren Verbindungen. Es sind diese Abhängigkeiten, die die Entwicklung entscheidend bestimmen und das Verständnis sich entwickelnder sozialer Netzwerke von anderen Disziplinen abzutrennen vermag. Ohne Abhängigkeiten zwischen Verbindungen besteht keine emergente Netzwerkstruktur [Brandes et al., 2013c].

Studien sozialer Netzwerke lassen sich funktional dadurch unterscheiden, ob das Netzwerk als erklärende oder abhängige Variable untersucht wird. Wir beschäftigen uns in dieser Arbeit primär mit letzterem Fall, also vermehrt mit Erklärungen für die Entwicklung von Netzwerken, und zum Beispiel weniger mit den Auswirkungen einer gegebenen Struktur auf den möglichen Einfluss einzelner Akteure.

In sozialen Netzwerken werden die (dyadischen) Variablen der Beziehungen jedoch häufig durch (monadische) Informationen über Akteursattribute erweitert. Aufgrund der wechselseitigen Wirkung von Akteursattributen auf die Entstehung von Kontakten (soziale Auswahl) und der Anpassung des individuellen Verhaltens an die gegebene Struktur (sozialer Einfluss) ist die Rolle des Netzwerkes sowohl als erklärende als auch als abhängige Variable immanent.

Der Einleitung dieser Arbeit folgt eine gründliche Darstellung von gegenwärtigen methodischen Ansätzen zur Modellierung, Analyse und Visualisierung sich entwickelnder sozialer Netzwerke. Im Fokus steht dabei die Granularität der verfügbaren zeitlichen Information innerhalb der beobachteten Daten. Die weiteren Kapitel vereinen vier wesentliche Beiträge und geben einen abschließenden Ausblick.

Kapitel 3 Wahrhaft strukturelle Analysen jenseits der isolierten oder (im statistischen Sinne) unabhängig aggregierten Betrachtung von dyadischen Beziehungen bilden den eigentlichen Kern der soziologischen Netzwerktheorie. Um dieser Theorie auch in der Praxis gerecht zu werden, plädieren wir für einen triadischen Ansatz zur Bestimmung der primären Akteursgruppen in sozialen Netzwerken. Eine entsprechende Methode zur Extraktion von "Simmelschen Netzwerksäulen" (*Simmelian backbones*) lässt sich quantitativ validieren und nutzt zudem fundamentale Eigenschaften sozialer Netzwerke für eine effiziente Berechnung.

Kapitel 4 Informationsvisualisierung ist ein wesentlicher Bestandteil für das bessere Verständnis sich entwickelnder sozialer Netzwerke. Die klassische Darstellung ist das Soziogramm und, ferner, die Soziomatrix. Eine große Herausforderung in der Visualisierung zeitabhängige Netzwerkdaten besteht in der kognitiv vertretbaren Repräsentation aufeinanderfolgender Zustände. Wir präsentieren einen völlig neuartigen Ansatz zur Darstellung von Zeitreihen asymmetrischer Netzwerkdaten, die "Gestaltmatrix". Diese Technik basiert auf der Kombination von Wortgrafiken (*sparklines*), mehrdimensionalen Daten-Glyphen und elementaren Gestaltgesetzen. Der Mehrwert solcher "Gestaltgrafiken" (*gestaltlines*) wird im Anhang dieser Arbeit zudem auch an nicht relationalen Daten aufgezeigt.

Kapitel 5 Ein fundamentales Problem für die effiziente und valide Bestimmung von wesentlichen Prinzipien in der Entwicklung sozialer Netzwerke betrifft die Frage der gegenseitigen Abhängigkeit einzelner Veränderungen. Wir definieren daher einen allgemeinen Rahmen für den systematischen Vergleich bedingter Unabhängigkeitsmodelle und allgemeinerer Modelle zur Analyse dynamischer Netzwerkdaten. Unsere Resultate bestätigen grundsätzlich die Notwendigkeit von Abhängigkeitsannahmen, legen allerdings auch nahe, dass bedingte Unabhängigkeitsannahmen unterschiedlich stark ins Gewicht fallen und zeigen zudem eine ins Auge fallende Diskrepanz zwischen verschiedenen etablierten Modellen.

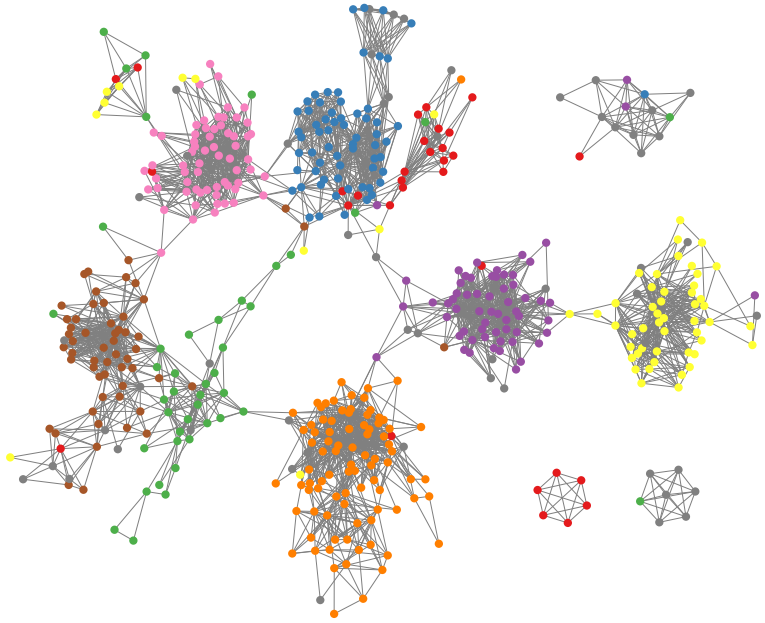
Kapitel 6 Die Vorhersage von neuen Beziehungen (*link prediction*) ist ein weiterer Faktor für das Verständnis sich entwickelnder Netzwerke. Nur wenige methodische Ansätze erlauben jedoch die genaue Betrachtung des Abstandes und der Reihenfolge vorangegangener Interaktionen; und Vorhersagen werden stattdessen auf zeitlich aggregierten Netzwerken getroffen. Demgegenüber adaptieren wir eine Methode des verteilten Rechnens, die Vektoruhr, und definieren eine intuitive und zugleich effiziente Variante für indirekte Informationsverbreitung in *sozialen* Netzwerken. Die Verwendung der daraus resultierenden Indikatoren steigert nachweislich die Treffsicherheit bisheriger Vorhersagemodelle.

Contents

Chapter 1.

Introduction

1



. . . "a set of points some of which are joined by lines" [Barnes, 1954] is a very intuitive picture for the concept of a social network. In the illustrative example above, the coloring provides anecdotal evidence for the efficacy of Simmelian backbones (cf. Chapter 3) w.r.t. revealing essential structure and primary actor groups.

1.1. Motivation: Understanding Evolving Social Networks

> *In all things which have a plurality of parts, and which are not a total aggregate but a whole of some sort distinct from the parts, there is some cause*
>
> *(Aristotle, Metaphysics)*

Social networks are constructs capturing interdependencies among seemingly autonomous social actors [Borgatti et al., 2009]. Any such complex web of interwoven social relationships neither appears out of the blue nor can it be expected to remain unchanged over the course of time. Quite the contrary. Ties are established, gain in strength, may slowly decay or terminate at once — all in mutual dependence. Social networks are fundamentally dynamic, ever evolving over time.

"Network science is itself more of an evolving network than a paradigm expanding from a big bang." [Brandes et al., 2013c, Claim 7] Presumably, a better understanding of evolving social networks – as they arise from various social processes – might be a key to explaining a wide range of social phenomena. As a matter of fact, the *social network* paradigm has become widely adopted in the social science, ever since Barnes [1954] coined the term by "trying to form an image for a multi-dimensional concept":

> The image I have is of a set of points some of which are joined by lines. The points of the image are people, or sometimes groups, and the lines indicate which people interact with each other. We can of course think of the whole of social life as generating a network of this kind. [Barnes, 1954]

This decidedly relational perspective – *the unit of analysis is the dyad* (a pair of actors) rather than a monad (a singleton actor) – has proven extremely convincing. Indeed, studies of social phenomena by means of network representations can be observed with (exponentially) increased regularity [Borgatti and Halgin, 2011].

Network structure is commonly described by the concept of a graph, i.e. as the sum of constituent dyadic relations. Still, social networks are not mere collections of dyadic variables. Rather, ties are systematically patterned and thus embedded beyond the dyadic level. Since dyads overlap at actors, the presence of ties is likely to conditionally depend on the presence of other ties; "The nature of networks leads to dependence between actors, and also to dependence between network ties." [Snijders, 2011] These interdependencies fundamentally drive network evolution and sets the understanding of evolving social networks apart from other disciplines; "Without dependence among ties, there is no emergent network structure." [Brandes et al., 2013c]

In network-based research, structure can play the role of both dependent and explanatory variable. Here, we are mainly interested in "why and how networks form the way they do" (i.e., *network formation*) rather than "why and how networks influence other outcomes" (i.e., *network effects*) [Brandes et al., 2013a]; see, e.g., our studies in [Brandes et al., 2011] and [Brandes et al., 2012b] for examples of the latter.

Yet, networks will often play the role of both antecedents and consequences at the same time. In particular, social networks are typically enriched with a collection of changeable actor attributes, termed "behavior", that co-evolve over time. For empirically observed associations between network structure and actor behavior, there are often competing explanations with opposite directions of causality. Longitudinal social network data, i.e., network data over time, are thus crucial to assess whether the social embedding of an actor influenced the actor's behavior (social influence), or whether an actor's behavior prompted a change of relations (social selection) — the role of networks as both antecedents and consequences is inherent.

The individual contributions in this thesis toward a better understanding of evolving social networks are delineated in the following section. In Chapter 2, moreover, we provide the big picture by spelling out in detail the type of data assumed and by reviewing related work on modeling and visualization techniques. In particular, we emphasize the crucial role of available granularity in temporal information – a single network observation, multiple network observations, or continuous dyadic observations – and thoroughly identify appropriate methods.

1.2. Overview of Main Contributions

In this thesis we focus on novel modeling, visualization, and analysis approaches toward a better understanding of evolving social networks. The four main contributions are as follows.

1.2.1. Sociologically Informed Extraction of Essential Structure

Structural analyses beyond degree and density measures bear immense potential, because dyadic relationships are assumed to depend on the structural environment in which they are embedded [Brandes et al., 2013c]. In particular, the understanding of evolving social networks is deeply linked to underlying sociological theories on network formation and human interaction — which, e.g., attach particular importance to triadic settings [Krackhardt, 1998, 1999].

In online social networks like Facebook, for example, interactions related to friendship, kinship, business, interests, and other relationships may all be represented as catchall "friendships." Because several relations are mingled into one, the resulting networks exhibit relatively high and uniform density. As a consequence, the variation in positional differences and local cohesion may be too small for reliable analysis with aggregate density-based approaches.

In contrast, we recently introduced a method to efficiently identify the *essential* relationships in networks representing social interactions [Nick et al., 2013]. Our method is based on a novel concept of triadic cohesion that is motivated by Simmel's concept of membership in social groups [Simmel, 1950a]. In Chapter 3 we demonstrate that our **Simmelian backbones** are capable of extracting structure from Facebook interaction networks that makes them easy to visualize and analyze.

1.2.2. Alternative Visual Exploration of Network Evolution

Although information visualization is not the primary focus of our investigations, we highly valuate the use of visualizations as a means of exploration, as well as a plausible way to communicate findings. Analytically informed network visualizations are indispensable to the understanding of evolving social networks.

We have been working on a novel technique of data representation, **gestaltlines**, that combines three powerful concepts from information visualization by composing *sparklines* of *gestalt theory*-informed *multivariate glyphs*. In interdisciplinary work with psychologists we were able to demonstrate that gestaltlines allow to reveal patterns, trends, and outliers in complex multivariate sequence data clearly and effectively within single word-sized diagrams [Brandes et al., 2013b]; cf. Appendix A.

For the particular case of longitudinal network data we proposed a matrix representation of gestaltlines, the **gestaltmatrix**, that specifically supports the exploration of evolving dyadic relations and persistent group structure [Brandes and Nick, 2011]. As we will demonstrate in Chapter 4, the resulting data-rich diagrams differ from common graphical representations of longitudinal social networks – animations and small multiples of intermediate states – in that they provide a compact holistic view of evolution. In addition to being a very intuitive way of viewing relationships, gestaltmatrices complement and inform the modeling and analysis of longitudinal network panel data — detecting evolving patterns and, in particular, exceptional actors, is of great importance in identifying factors that govern the evolution of dyads, because current models rely on fairly strong homogeneity assumptions [Snijders et al., 2010b].

1.2.3. Justification for Complex Dynamic Network Models

Are tie changes in a social network independent, given an initial observation? Answering this question for given data is of high practical relevance since, if the conditional independence assumption is valid, network evolution can be modeled with simple and computationally efficient statistical techniques for independent observations.

We proposed a **lagged network statistic framework** that allows to systematically compare conditional independence models with more general models that are specifically designed for social network data [Lerner et al., 2013]. For instance, compare different change scenarios that ultimately result in a transitive triangle  configuration: a conditional independence model could not recognize an evolution starting from an initial single tie  as transitivity effect, since, e.g., a third tie closing a directed two-path  that was not present in the beginning, can only be taken into account by more complicated conditional dependence models.

As we will demonstrate in Chapter 5, we found that conditional independence models are inappropriate as a general model for network evolution and can lead to distorted substantive findings on structural network effects, such as transitivity. On the other hand, our results suggest that the conditional independence assumption becomes less severe when the time span between subsequent network observations is relatively short.

1.2.4. Exploitation of Detailed Timestamps in Relational Data

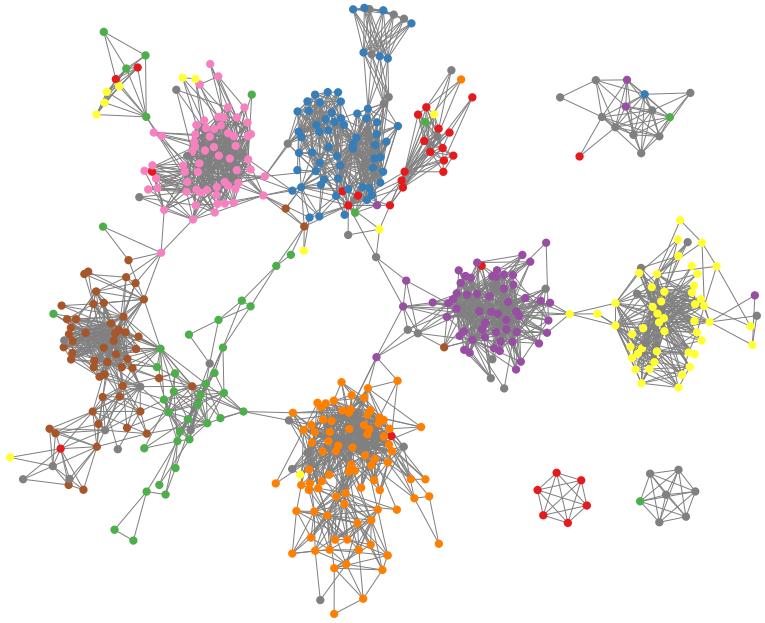
With an increasing amount of data on every aspect of daily activities, network information is often generated as a fine-grained sequence of individual time-stamped relational *events*, in which the exact minute or second of social interaction is known. In contrast, the most common approaches for analyzing evolving social networks assume network panel data, i.e. relatively coarse-grained temporal information on a sequence of relational *states*. As a consequence, many state-of-the-art techniques require one to first aggregate network event data into (time-sliced) static graphs — a step in which the ordering and spacing of events is lost.

Developing novel techniques that are able to deal with detailed timestamps in relational data will advance the understanding of evolving social networks [Butts, 2008, Brandes et al., 2009c], not least because an additional temporal dimension is introduced through processes such as viral marketing that take place on the network, rather than forming it. As one possible step in that direction, we build upon the work of Kossinets et al. [2008] and introduce a parameterized version of **social vector clocks** in a way which makes the vector clock concept both more scalable and more appropriate for social networks [Lee et al., 2013]. In a first application, which will be presented in Chapter 6, our experiments on link prediction [Liben-Nowell and Kleinberg, 2003] suggest that by taking into account the order and spacing of interactions, social vector clocks exploit different aspects of link formation so that their combination with previous approaches yields the most accurate predictor to date.

Chapter 2.

Social Networks: Data and Methods

2



Here we combine and extend background information on related work as originally provided in our papers underlying Chapter 3–6. This chapter therefore contains various text passages from Nick et al. [2013], Brandes and Nick [2011], Lerner et al. [2013] and Lee et al. [2013].

EMPIRICAL social networks undergo gradual change. Therefore, the granularity of available temporal information in social network data is a decisive factor in their understanding. On one hand, for instance, statistical models for longitudinal network data are less complicated than appropriate models for cross-sectional data, "because dependencies are spread out in time." [Snijders, 2011]. With regard to network visualizations, on the other hand, additional temporal information increase the level of complexity significantly [Moody et al., 2005].

Initially, in this section, we delineate the scope of our research toward a better understanding of evolving social networks by clarifying basic terminology and spelling out in detail appropriate methods for different type of data assumed in our investigations, i.e. cross-sectional *static data* (a single network observation), longitudinal *panel data* (multiple network observations), and time-stamped dyadic *event data* (continuous network observation).

2.1. Single Observation of Network Evolution

As we have outlined in the introduction, it is the relational perspective which is the distinctive feature of empirical social network studies; the unit of analysis is the dyad (pairs of actors) rather than the monad (singleton actors), and dyads overlap by design.

The most common type of data studied consists of a set of actors such as individuals or organizations, and – most important – one or more types of relations between them. Examples are friendship networks among pupils and the trading of goods between nations. "Ties are data on dyads" [Hennig et al., 2012] and typically modeled by a graph $G = (V, E)$, consisting of a set of vertices V representing the actors and a set of edges $E \subset V \times V \setminus \{(v, v) : v \in V\}$ representing the ties. Alternatively, an actor-by-actor table in form of a (binary) adjacency matrix $y = (y_{ij})_{i,j}$ with $y_{ii} = 0$ for all actor indices i, is used to indicate the existence ($y_{ij} = 1$) or non-existence ($y_{ij} = 0$) of a tie — both representations reveal the distinctively composite and overlapping nature of dyadic indices in network data. Relations may be symmetric (as is often assumed in friendship networks), i.e. $(u, v) \in E \implies (v, u) \in E$ and $y_{ij} = y_{ji}$, respectively, or asymmetric (as is typically the case in the trading of goods).

Besides dyadic information on the network structure, social networks are typically enriched with additional attributes of actors (monadic attributes) and ties (dyadic attributes). In particular, dyadic relations are often weighted, i.e., a numerical attribute value associated with each tie, $y_{ij} > 0$, indicates a specific quality of the relation; such as the volume of trade in the above example. The number and type of actor attributes is much more application-specific, and especially relevant for longitudinal network studies (as we will discuss in Section 2.2). For now, we assume a single observation of network evolution, i.e. the lack of temporal information whatsoever.

See Marsden [1990], Morris [2004], Handcock and Gile [2010], and Hennig et al. [2012] for in-depth discussions on social network data and data collection.

2.1.1. Example: Friendship Relations on Facebook

As a basic, yet rich, example for cross-sectional (static) network data, we introduce the `Facebook100` dataset. These networks were introduced and analyzed in Traud et al. [2011] and Traud et al. [2012]. The dataset contains 100 individual (single) snapshots of friendship relationships among *all* Facebook user profiles at distinct American academic institutions in September 2005; the networks range in size from 769 nodes and 16,656 edges (California Institute of Technology; Caltech) to 36,371 nodes and 1,590,655 edges (presumably the University of Texas at Austin).

The data has many desirable characteristics. For example, it is not sampled,[1] and it comes from a service which at the time of data collection was widely and intensively used by students. Furthermore, the dataset includes multiple actor attributes that were extracted from the corresponding Facebook profiles, namely, each student's `gender`, `year` of graduation, `major` and `minor` fields of study, attended `high-school`, and `dorm`itory residence on campus, as well as a distinction of student or faculty `status`. Not every user filled in every field of the profile, and so for all attributes except for student/faculty status many values are missing.

As Snijders [2011] has noted, "More scientific progress can be made when data are available for several networks that may be regarded, in some sense, as replications of each other"; in this regard, the `Facebook100` dataset is certainly a valuable collection of cross-sectional network data. We make particular use of the data in Chapter 3.

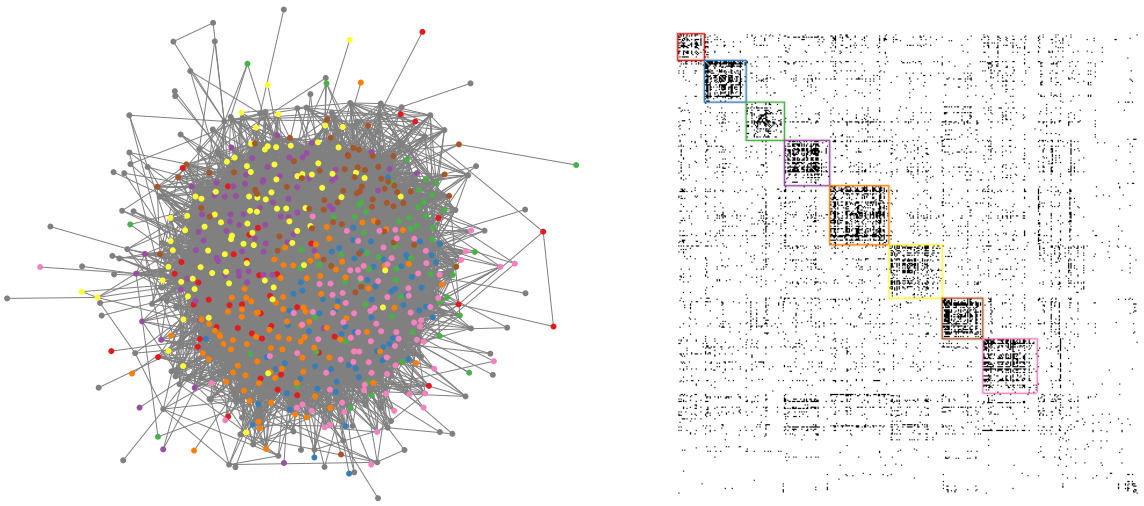
2.1.2. Sociogram and Sociomatrix

"Visualization has been instrumental in the study of social networks from the very beginning. [...] The two main purposes of network visualizations are exploration of data and communication of findings." [Brandes et al., 2013a]

The most frequent form of graphical representation for cross-sectional social network data is the *sociogram*, a "node-link" representation which directly corresponds to the image of points and lines mentioned in the introduction; cf. Figure 2.1(a). A second, less common representation is the *sociomatrix*, a tabular representation of dyadic information corresponding to the adjacency matrix of a graph; cf. Figure 2.1(b) and see Forsyth and Katz [1946] for one of the earliest examples. Both terms, sociogram and sociomatrix, have been introduced in Moreno [1953]. Recently, in combination with specific interaction concepts, matrix representations have re-gained some popularity in systems for exploring large network data such as Elmqvist et al. [2008].

The crucial algorithmic challenges for node-link diagrams are placement of nodes and routing of edges. Such layout problems are the main focus in graph drawing [Battista et al., 1999, Kaufmann and Wagner, 2001], where many fundamental techniques have been developed that can be adapted for specific scenarios. Exemplary designs are proposed, e.g., in Brandes et al. [2001], Brandes et al. [2003], and Perer and Shneiderman [2006]. A sub-category of node-link diagrams is formed by attribute-based

[1] According to Traud et al. [2012], the data "was sent directly to us in anonymized form by Adam D'Angelo of Facebook" (who was Facebook's CTO at that time).



(a) sociogram (b) sociomatrix

Figure 2.1.: The most frequent forms of graphical representation for cross-sectional social network data, illustrated on the Caltech network in the `Facebook100` data. According to the guideline provided in the main text, a sociomatrix can be more suitable for highlighting higher level organization than a sociogram if the network is rather dense; used colors correspond to available dorm attribute information.

designs as exemplified in PivotGraphs [Wattenberg, 2006].

Similarly, the main algorithmic challenge for matrix-based network representations is the ordering of actors which determines the permutation of rows and columns [Díaz et al., 2002]. In social network analysis, orderings are often determined to highlight higher level organization in the matrix. The corresponding technique of blockmodeling refers to substantively meaningful rearrangements that (visually) reveal regularities of the network structure within the matrix cells. For instance, one may be interested in cohesive groups and thus order the matrix so that locally dense groups form blocks along the diagonal. While this is not the only criterion, it is certainly among the one most commonly used. For most scenarios, however, finding an optimal permutation is $\mathcal{NP}$-hard. Various heuristics to calculate acceptable solutions for given criteria have been proposed. A comprehensive overview of blockmodeling techniques is provided in Doreian et al. [2005].

Note that a sociomatrix representation needs quadratic space in the number of actors, but avoids occlusion problems resulting from overlapping actors and ties in a sociogram. Based on discussions such as Ghoniem et al. [2005], Hennig et al. [2012] therefore conclude that "Sociograms are more suitable for sparse networks and the investigation of indirect linkages" and "Sociomatrices are more suitable for dense networks and the investigation of partition blocks." A hybrid approach combining node-link diagrams and matrix representations for dense subgraphs is presented in Henry et al. [2007].

2.1.3. Exponential Random Graph Models (ERGMs)

Besides network visualization, the methodological toolbox for empirical social network studies can be organized into three main compartments, namely, indexing: "The assignment of values to predetermined substructures of any size", grouping: "The identification of substructures and membership in them", and modeling: "The use of statistical models for assessment and inference." [Brandes et al., 2013a]

Since the primary focus in understanding evolving social networks is network formation (i.e. networks as dependent variables) rather than network effects (i.e. networks as explanatory variables), we will concentrate on modeling approaches in the following. Extensive reviews on indexing and grouping can be found, e.g., in Wasserman and Faust [1994] and Brandes and Erlebach [2005]. A more detailed review on statistical analysis and modeling of cross-sectional network data than we provide here, can be found, e.g., in Kolaczyk [2009].

Less realistic social network models. As we have noted before, one of the most universal characteristics of network data is the property that dyadic observations are typically not independent. For instance, if two actors have a common friend, the probability of a friendship tie among them is often highly increased. Likewise, actors often show a tendency to connect with those that are already well connected in the network. Such non-independence among observations – which is typically the actual substantive interest in analyzing network data – can make the design of valid network models a challenging task.

In turn, network models assuming independence of tie observations disqualify from being appropriate in general. In the traditional $\mathcal{G}(n, p)$ random graph model [Gilbert, 1959], which is a slight modification of Erdős and Rényi [1959], for instance, the number of nodes n is fixed and each dyad gives rise to an edge with probability $0 < p < 1$ (independent and identically distributed); see Bollobás [2001] for an extensive review on classical random graph theory. Similarly, planted partition models, such as those being used in Brandes et al. [2009b] use different (independent) tie probabilities within and between prespecified groups and thus allow for varying (expected) densities in different (predefined) regions, but do not incorporate edge dependencies at all.

Other examples of non-realistic *social* network models include those approaches that are crafted to reproduce some (but insufficiently many) structural characteristics that have been empirically observed in real-world networks. Examples include the preferential attachment model of Barabási and Albert [1999] to obtain power law degree distributions, as well as "small worlds" [Milgram, 1967], i.e. the generation of sparse networks with high local density and short average distances between nodes [Watts and Strogatz, 1998].

More realistic social network models. In contrast to the aforesaid approaches, Frank and Strauss [1986] proposed a quite general class of models, *Markov Random Graphs*, in which a dyadic observation might conditionally depend on the presence or absence of other *incident* ties (but dyads that do not have a node in common are always conditionally independent). Such local (Markov) properties correspond to global (Gibbs) distributions that can be deduced from the Hammersley-Clifford theorem [Besag, 1974]. In particular, it turns out that Markov random graphs are a subclass of exponential-family random graph models (ERGMs), which have become the predominant approach to model cross-sectional social network data — see, e.g., Robins et al. [2007a], Robins et al. [2007b], and Lusher et al. [2013] for definitive introductions.

A random graph model on a fixed set of vertices belongs to the class of exponential (family) random graph models[2] if the probability distribution for (directed or undirected) networks – i.e. the joint distribution of ties – is specified by

$$P(Y = y) = \frac{1}{Z(\theta)} \cdot \exp\left(\sum_{\ell=1}^{k} \theta_\ell \cdot s_\ell(y)\right) \quad , \tag{2.1}$$

where the s_ℓ are statistics mapping a network y to a real number, the $\theta_\ell \in \mathbb{R}$ are associated parameters modeling the influence of the statistics on the probability of the network y, and $Z(\theta)$ is a normalization constant ensuring that the probabilities of all possible networks on the fixed set of actors sum up to one. Appropriate statistics introduce dependence among dyads. For instance, specifying an ERGM with statistics that count the number of edges, k-stars[3] ($k = 2, \ldots, n-2$) and triangles in the network corresponds to specifying a (homogeneous) Markov random graph.[4]

Moreover, the Hammersley-Clifford theorem implies that every random graph model on a fixed set of vertices which assigns non-zero probabilities to all networks can be described as exponential random graph model. The $G(n, p)$ model with $0 < p < 1$, for instance, corresponds to an ERGM that is specified with a single statistic counting the number of edges in the network and an associated parameter $\theta = \log \frac{p}{p-1}$ according to the log-odds for a tie being present[5] — clearly, a single statistic that is only controlling for density does not introduce dependence among dyads.

[2] Now and again, but less and less often, referred to as p^* models [Wasserman and Pattison, 1996].

[3] A k-star is a network structure with one node of degree k that is connected to k nodes of degree 1. In this sense, an edge can also be interpreted as forming a 1-star.

[4] Note that among these statistics the triangle statistics is special, and complementing the (implicit) modeling of density and degree distribution.

[5] If more than one statistic is involved in the ERGM specification, the log-odds interpretation is less straightforward: for a given statistic, then, the associated parameter can be interpreted as (conditional) log-odds for a tie being present, if that tie is increasing that statistic by one, but leaves all other statistics unchanged, i.e. *all else being equal*.

With the ERGM framework, as we will elaborate in the following, it is possible to 1) propose, 2) fit, and 3) evaluate (simplified) assumptions about which mechanisms are at play and lead to empirically observed network features.

First, we can select any set of network statistics to specify a parameterized family of probability distributions and, in this way, model some constituents that are considered necessary in understanding an emergent network structure. In particular, statistics can not only be selected to model endogenous (structure-based) effects, but also incorporate exogenous (attribute-based) effects which, e.g., might give rise to "homophily", i.e. a predominant tendency to join with similar others. [Lazarsfeld and Merton, 1954, McPherson et al., 2001]

Second, to estimate the influence of these effects, we look for associated parameter values such that the observed network is rather typical than atypical. In other words, we look for a particular member of the parameterized family for which the observed network has rather high probability. In theory, we would select those parameters that fit best to the observed data (maximum likelihood estimation). In practice, direct computation of probabilities in exponential random graph models is computationally intractable for all but the smallest networks, since the number of networks and, hence, summation terms for determining the normalization constant $Z(\theta)$ grows exponentially with network size. Sophisticated stochastic approximation techniques exist [Snijders, 2002a] but are still computationally expensive, often lack reliable confidence intervals, and might suffer from model degeneracy (see below) — therefore, to date, ERGMs have been applied to networks with only hundreds rather than tens of thousands of nodes.

Third, since the best fit is not necessarily a good fit [Hunter et al., 2008], it is current practice to compare sample graphs from the estimated probability distribution with the observed network by means of a (different) set of network statistics.[6] Sanity checks are especially important, since ERGMs are prone to (near) model degeneracy, such as bimodal probability distributions of extremely sparse and extremely dense networks [Handcock, 2003]. To overcome the latter and improve the fit to empirical data, various authors have proposed new specifications of (curved) exponential family random graph models recently [Snijders et al., 2006, Robins et al., 2009]. For example, the geometrically weighted edgewise shared partner (`gwesp`) statistic can be used to control for the marginal effect of additionally closed triangles [Hunter and Handcock, 2006].

In summary, exponential random graph models provide a powerful framework for statistically analyzing cross-sectional data of evolving social networks, but a substantive specification of relevant statistics and the inference of associated parameters is far from trivial. For an application of ERGMs to the `Facebook100` dataset refer to Traud et al. [2012].

[6] Note that both approximation and model fit techniques involve the drawing of samples from the specified ERGM. This can be realized with a Gibbs sampling strategy that avoids the calculation of the normalization constant [Snijders, 2002a].

2.2. Multiple Observations of Network Evolution

The formation of social networks is a continuous process; time is what makes up the structure. Yet, cross-sectional network data only provide a single snapshot of that persistent evolution. The picture becomes very different if longitudinal data become available, since this entails explicit information on gradual changes in the network structure. As Snijders [2011] has noted, "Modeling network dynamics is less complicated than modeling single network observations because dependencies are spread out in time."

In particular, dynamic network data (in any shape) is essential in understanding causality in the co-evolution of network ties and changeable actor attributes; cf. Figure 2.2. As we will elaborate below, longitudinal models like the one discussed in Steglich et al. [2010] aim to disentangle, e.g., a propensity toward homophily (social selection) [McPherson et al., 2001] from contagion mechanisms (social influence) [Friedkin, 1998] — although there is ongoing discussion to which degree this is possible at all [Shalizi and C.Thomas, 2011].[7]

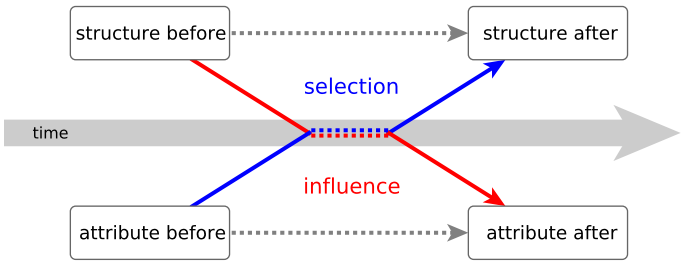


Figure 2.2.: The entangled mechanisms of social selection and social influence and its impact on the co-evolution of network structure and behavioral attributes, possibly with feedback. "Disentangling the effects of selection and influence is one of social science's greatest unsolved puzzles." [Lewis et al., 2012] Ignoring some of these dependencies, as well as additional environmental factors, may lead to questionable conclusions [Christakis and Fowler, 2007, Cohen-Cole and Fletcher, 2008a,b].

The basic scenario of longitudinal network data has been delineated, e.g., in Snijders et al. [2010b]: "We assume here that the empirical data consist of two, but preferably more, repeated observations of a social network on a given set of actors; one could call this *network panel data*." (A different form of dynamic network data, *dyadic event data*, will be treated in Section 2.3.)

Waves of consecutive network observations naturally arise from data collection strategies in empirical social network studies. That is, often based on surveys, a

[7] As Snijders has noted elsewhere: "*Disentangling* selection and influence is possible only under the assumption that the available observed networks and individual variables contain all the variables that play a role in the causal process, and if moreover a number of distributional assumptions are made" [Citation form the SOCNET mailing list; the message is available as archive entry at `http://www.lists.ufl.edu/cgi-bin/wa?A2=ind1106&L=SOCNET&P=R11428`]

dynamically evolving network of relational states (such as friendship) is monitored at two or more discrete points in time. Data collected in this way (i. e., the *observation*) corresponds to a number $T \geq 2$ of binary matrices $y^{(1)}, \ldots, y^{(T)}$ of order $n \times n$, where n, the number of actors, is kept constant over time. Typical values for T are in the range of 2–30 observations, with a bias toward the lower end. Moreover, the majority of empirical studies deals with networks of 10–100 actors.

Additional attribute information within empirical social network data is particularly valuable in the case of longitudinal studies, since structural changes in the network might not only depend on the structure itself (e.g., friends of friends become friends) but, among other, can also depend on *actor-level covariates* (such as favorite instrument) and *dyad-level covariates* (such as having played in the same band). In this context, changeable actor-level attributes are often referred to as "behavior" and, as we have indicated in Figure 2.2 above, in the same way that actor behavior might change network structure (social selection), the network structure might prompt a change in actor behavior (social influence).

2.2.1. Example: Newcomb Fraternity and Knecht Classroom Data

As illustrative examples for the type of data that is typical for longitudinal network panel studies we introduce two datasets. The two datasets differ, among others, in the availability of covariates and in the length of the time interval between observation points.

The first dataset, well-known as *Newcomb (fraternity) data* [Nordlie, 1958, Newcomb, 1961], encodes 15 weekly snapshots of an evolving social network among 17 previously unacquainted male students attending the University of Michigan in the fall of 1956. The original data comprises complete sociometric preference rankings, 'like best' to 'like least', but does not include any actor-level or dyad-level covariates. During the study all men got free accommodation in fraternity housing, hence the name.

The development of friendships in the Newcomb fraternity data has been (re)analyzed numerous times, e.g., in Trappmann et al. [2011], Moody et al. [2005], Doreian et al. [1996], Nakao and Romney [1993], and White et al. [1976]; consequently, there is no doubt that it presents a case in point. Often, the data has been dichotomized by only keeping the top-k nominations for each actor without distinction of their ordering. The transformed data, then, encodes the hypothetical answers to a typical question in other empirical social network studies such as 'name your k best friends', 'indicate your k most important business partners', etc. Note that by this transformation the outdegrees (i.e., votes cast) are constant over the actors and over the different time steps and thus can not be consulted to explain network evolution. We will make extensive use of the Newcomb fraternity data in Chapter 4 and Chapter 5.

The second illustrative dataset, subsequently referenced as *Knecht (classroom) data*, is a subset of the data collected by Knecht [2008]. The data is about a friendship network in a Dutch school class. Snapshots of the evolving social network have been collected less frequently, at four points in time with intervals of three months, but

contain additional covariate information. Among others, actor-level covariate data about the gender of the pupils and a dyadic covariate encoding whether two pupils were in the same primary school are available. "Network data were assessed by asking students to indicate up to 12 classmates which they considered good friends. The average number of nominated classmates ranged between 3.6 and 5.7 over the four waves, showing a moderate increase over time." [Snijders et al., 2010b] We will make use of the indicated friendship data and constant covariate information in Chapter 5.

In Section 2.2.3 we will discuss the stochastic actor-oriented model framework which is the predominant approach to model and analyze empirical network panel data. Note that both Newcomb fraternity data and Knecht classroom data have been used as illustrating case studies on how to do so; refer to Snijders [1996] and [Snijders et al., 2010b], respectively.

2.2.2. Time-Varying Network Visualization

The most common graphical representations of time-ordered network sequences appear to be animations and small multiples of cross-sectional network views, depending on the type of media available.

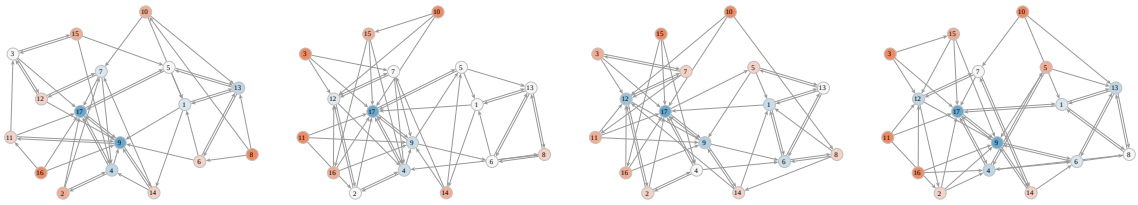
For node-link representations, there are at least two (often competing) optimization criteria at work: firstly, to define high-quality layouts of individual intermediate states (layout quality) and, secondly, to ensure smooth transitions between individual drawings (layout stability). That is, the basic layout problem is made more severe by additional coherence constraints that are meant to alleviate the difficulty of retaining a mental map of the network's structure [Eades et al., 1991, Misue et al., 1995, Archambault et al., 2011].

In the *online* scenario, the layout of each network in the sequence is solely based on past and the present structure, while in the *offline* scenario a layout algorithm can also incorporate knowledge on future network states. Brandes et al. [2012a] have summarized three techniques that are commonly applied, namely, *aggregation*: "All graphs in the sequence are aggregated into a single graph that has one vertex for each actor. The position of each individual vertex instance in the sequence is determined from a layout of the aggregated graph.", *linking*: "All graphs in the sequence are combined into a single graph that has one vertex for each occurrence of an actor, and an edge is created between vertices representing the same actor in consecutive graphs. A layout of this graph directly yields positions for all vertex instances in the sequence.", and *anchoring*: "Using auxiliary edges, vertices are connected to immobile copies fixed to a desired location which may be, for instance, the previous position in an online scenario, or a reference position in an offline scenario."

With regard to the (physical) presentation of (logically) determined consecutive layouts, the use of animation is a straightforward solution for incorporating available temporal information. However, a mapping of empirical time to display time requires special media and thus is not suitable for traditional print publication. An alternative to plain animation is interaction, in which time and focus are chosen explicitly — see, e.g., Yi et al. [2010] for a matrix-based interaction system and refer to Tikhonova

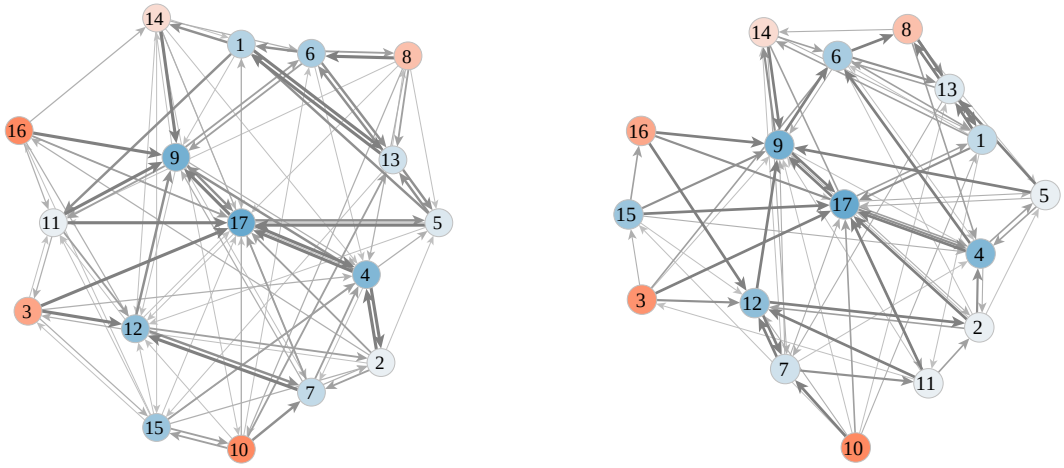
et al. [2010] for a systematic treatment of this approach.

Alternatively, if dynamic media are not available or simultaneous cross-time comparisons are of importance, snapshots of an animation are often displayed in small multiples [Tufte, 1990]. The example in Figure 2.3 is based on an aggregation approach (cf. discussion above) over all time points in the Newcomb fraternity data. Such aggregate networks can be used for representation themselves as in Figure 2.4.



(a) wave 2 (b) wave 6 (c) wave 10 (d) wave 14

Figure 2.3.: Small multiples of top 3 friendship nominations in the Newcomb fraternity data. Red/blue color scale represents lower/higher-than-expected popularity.



(a) first five weeks (b) last five weeks

Figure 2.4.: Example for small multiples of aggregate summary views on network panel data. Link prominence according to cumulated top 3 friendship nominations in the Newcomb fraternity data. Node coloring according to overall popularity.

Both animation and small-multiples are treated in depth in Brandes et al. [2012a]. Even with the most sophisticated techniques, though, attempting to describe and compare the evolution of multiple relationships or detecting interesting patterns – such as non-requited friendship nominations – often results in cognitive overload. In Chapter 4, therefore, we present a completely novel approach for representing evolving dyadic relationships holistically.

2.2.3. Temporal ERGMs and Stochastic Actor-Oriented Models

Traditional approaches to modeling network formation focus on (re)producing but one (static) network, although occasionally the model process incorporates some notion of growth [Newman, 2003]. In the preferential attachment model of Barabási and Albert [1999], for instance, nodes are subsequently added to the network and preferably connected to popular nodes.[8]

In contrast, only recently, longitudinal network models have been proposed that explicitly take into account the transition from one observed network state into another, i.e. focus on (re)producing (dynamic) evolution. The ultimate aim in studying such network dynamics is to "shed light on the underlying theoretical micro mechanisms that induce the evolution of social network structures on the macro level." [Snijders et al., 2010b] We will treat the most commonly used frameworks below. The first is a temporal version of exponential random graph models, the second is the class of so-called stochastic actor-oriented models.

Formally, given longitudinal network data, the observed adjacency matrices are assumed to be realizations of a continuous-time random process $Y^{(t)} \in \{0,1\}^{n\times n}$, $t \in \mathbb{R}$, at observation times $t_1 < \cdots < t_T$. The observation that the i,j-th entry of y equals one at time t_h (which we denote by $y_{i,j}^{(h)} = 1$) encodes that there is a tie from actor i to actor j at that time; else $y_{i,j}^{(h)} = 0$. The network at a particular observation point is assumed to result from the preceding one and a sequence of unobserved tie-change events.[9]

Temporal ERGMs. Bringing in a temporal component into the ERGM class [Hanneke et al., 2010, Cranmer and Desmarais, 2011] is straightforward and closely related to the concept of lagged network statistics that we will introduce in Section 5.1.1. More specifically, given an observed sequence of networks $y^{(1)}, \ldots, y^{(T)}$, the conditional probability $P(Y^{(t_h)} = y^{(h)} | y^{(h-1)})$ is modeled as in Eq. (2.1) where we allow the statistics to be functions of both networks $y^{(h)}$ and $y^{(h-1)}$. In formulas, for all $h \geq 2$

$$P(Y^{(t_h)} = y^{(h)} | y^{(h-1)}) = \frac{1}{Z(\theta^{(h)})} \cdot \exp\left(\sum_{\ell=1}^{k} \theta_\ell^{(h)} \cdot s_\ell(y^{(h)}, y^{(h-1)})\right) \quad . \tag{2.2}$$

Thus, statistics might purely depend on the network $y^{(h)}$ (expressing structural patterns at time t_h without a temporal component) or on both the network $y^{(h)}$ and $y^{(h-1)}$ (expressing dependence on the preceding observation).[10] Note that statistics depending only on the previous network $y^{(h-1)}$ would lead to a non-identifiable model since their value is the same for all networks at time t_h.

[8] In the broader literature on machine learning a similar concept is known as Indian Buffet Process [Griffiths and Ghahramani, 2011].

[9] For keeping notation concise, we will not explicitly write dependence on covariates in the following.

[10] We make the (Markovian) assumption that the network evolution in the interval (t_{h-1}, t_h) is stochastically determined by the network $y^{(h-1)}$ but conditionally independent of the previous networks $y^{(1)}, \ldots, y^{(h-2)}$, i.e., $\forall h \geq 2 \;\; P(Y^{(t_h)} = y^{(h)} | y^{(h-1)}) = P(Y^{(t_h)} = y^{(h)} | y^{(1)}, \ldots, y^{(h-1)})$.

Stochastic Actor-Oriented Models. Another class of network models we have not touched upon yet are strategic interaction models, which are mainly rooted in algorithmic game theory and economics [Jackson, 2008, Goyal, 2009]. As opposed to (global) probability distributions in random graph models based on empirically observed characteristics ("how does the network form?"), strategic interaction models take a (local) perspective and produce equilibrium networks that are stable regarding the underlying incentives of involved actors ("why does the network form?"). Common scenarios are based on the tradeoff between costs for creating network ties, and actor (agent) incentives (benefits, rewards) to do so — such as reducing the average distance to all others [Fabrikant et al., 2003]. Yet, such equilibrium networks are often highly crafted; see, e.g., our analysis in Brandes et al. [2008].

Attempts to combine random graph models with strategic interaction models[11] and carrying them over on dynamic networks have remained largely untouched so far, with the exception of so-called stochastic actor-oriented models (SAOMs).

SAOMs [Snijders, 2001, 2005] assume that the observed networks $y^{(1)}, \ldots, y^{(T)}$ are snapshots of an underlying dynamic process driven by myopic actor decisions. This process is modeled separately for each interval $[t_{h-1}, t_h]$, $2 \leq h < T$, as a stochastic chain of dyad flips (creating a new tie or dissolving an existing one) leading from $y^{(h-1)}$ to $y^{(h)}$. Changes do not occur synchronously, but successively (with arbitrarily small time steps) for single relations. At each moment $t_{h-1} + \Delta t$, $0 \leq \Delta t < t_h - t_{h-1}$, dyad-change probabilities depend only on the current network structure y. Network evolution is thus modeled as a continuous-time Markov process.

The models are actor-oriented in the sense that the dyad flips are assumed to be performed by the actors, with each actor i, $1 \leq i \leq n$, controlling only his/her outgoing (binary) relations y_{ij} to the $n-1$ other actors. That is, when actor i gets the opportunity to change the current network y, he/she randomly chooses to flip dyad y_{ij} into its opposite $1-y_{ij}$ with a probability that is deduced from the enhancement of his/her position in the resulting network. The enhancement of i's position is measured by the *evaluation function* (or "objective function")

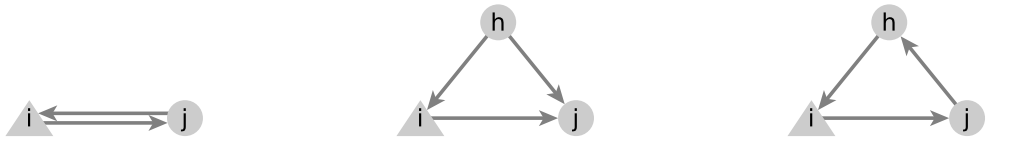
$$f_i(\theta^{(h)}, y, y^{\neg ij}) = \sum_{\ell=1}^{k} \theta_\ell^{(h)} \left(s_{i\ell}(y^{\neg ij}) - s_{i\ell}(y) \right), \tag{2.3}$$

where $y^{\neg ij}$ refers to flipping a dyad into its opposite and statistics $s_{i\ell}$ count certain *local* network configurations such as the number of reciprocated ties, the number of mediating ties, or the number of ties within 3-cycles *that actor i holds*; cf. Figure 2.5. Assuming that the focal actor is uniformly chosen, the probability that y_{ij} will be flipped in the next step is given by

$$n^{-1} \cdot \frac{\exp(f_i(\theta^{(h)}, y, y^{\neg ij}))}{\sum_{k=1}^{n} \exp(f_i(\theta^{(h)}, y, y^{\neg ik}))} \quad .^{12} \tag{2.4}$$

[11] Interesting work in that direction has been recently presented by Viviana Amati and Ulrik Brandes: "On ERGMs as the Outcome of Network Formation Games" (Sunbelt 2012), "Interpreting Near-degeneracy of ERGMs in Terms of Socially Desirable Equilibria" (Sunbelt 2013); unpublished.

[12] These change probabilities correspond to a discrete (multinomial) choice model in which the ran-



(reciprocated ties) (mediating ties) (3-cycles)

$$s_{i1}(y) = \sum_j y_{ij} y_{ji} \qquad s_{i2}(y) = \sum_{j,h} y_{ij} y_{hi} y_{hj} \qquad s_{i3}(y) = \sum_{j,h} y_{ij} y_{jh} y_{hi}$$

Figure 2.5.: Illustrative examples of three local network statistics that could be incorporated in the evaluation function of the stochastic actor-oriented model; a detailed description of various theoretical micro effects is provided in Ripley et al. [2012].

The network evolution is thus modeled as the outcome of micro-rules (agency) that lead to results at the macro-level (structure). Note that not changing any tie (denoted by $y^{\neg ii}$) is also an option for actor i. Moreover, the model also allows to define *structural zeros* (i.e., non-possible connections).

As we have noted above, the SAOM framework also provides a mean to investigate the co-evolution of network structure and behavioral actor attributes. In this case, actors do not only control their outgoing ties, but also their behavioral attributes; i.e. a focal actor i gets the opportunity to flip a dyad or alternatively may change his/her behavior. Behavioral attributes are assumed to be coded as ordinal "levels", and with each change event can be incremented or decreased by one unit (or remain unchanged). Both types of change, structural and behavioral, are modeled with individual evaluation functions: Social selection is modeled with the specification of a *structural evaluation function* (including attribute-based network effects) and, vice versa, social influence is modeled with the specification of a *behavioral evaluation function* (including structural-based network effects).[13]

Having specified appropriate evaluation functions, the associated parameters can again be estimated to fit the observed data [Snijders et al., 2010a, Snijders, 2005] and, ideally, indicate statistically significant (micro) effects on the network evolution (as observed at the macro level). Complementing network visualizations to assess the fit of actor-based models have been proposed in Brandes et al. [2012a, Section 5].

Since empirical researchers can draw upon a range of corresponding software implementations, such as the RSiena package [Ripley et al., 2012],[14] the SAOM is in wide use and diverse applications have been published recently. However, the underlying procedures are based on expensive Markov chain Monte Carlo simulations for approximating a method of moments estimator [Snijders, 2002b] or the maximum likelihood

dom term in the utility function is modeled as an independent and identically distributed extreme value distribution, more precisely, the Gumbel distribution [McFadden, 1973].

[13] The evaluation function for the behavioral changes typically includes a basic linear effect and (for non-dichotomous attributes) also a quadratic shape effects to model self-correcting and self-reinforcing mechanisms. [Snijders et al., 2010b]

[14] See `http://www.stats.ox.ac.uk/~snijders/siena/` for an overview of available programs; SIENA stands for Simulation Investigation for Empirical Network Analysis.

estimator [Snijders et al., 2010a], and therefore restrict the practical applicability to network sizes of about 1000 nodes; see Lewis et al. [2012] for an application to Facebook friendship relationships with $n = 1001$ nodes (and $T = 4$ time-steps). By design, moreover, the underlying discrete choice model implies that actors in the SAOM have perfect knowledge of the complete network. If anything, this assumption can only be justified for small networks.

Note that temporal exponential random graph models and stochastic actor-oriented models make quite different assumptions. ERGMs assume that the network is in an equilibrium distribution. Although the structure of this equilibrium might be shaped by the previous network, it is totally ignored how many tie-change events are needed to go from one observed network to the next. SAOMs, on the other hand, assume that there is an unobserved chain of events in which exactly one tie changes. In this framework there is no room for higher-order changes of mutually dependent tie-change events, e. g., events where three people meet each other in one single point in time or one actor synchronously cuts friendship to a group of actors.

At the core of both (t)ERGM and SAOM, however, is the linear combination of explanatory statistics that are assumed relevant in understanding evolving social networks and, indeed, an interesting relation between both model frameworks is the following: if the focal actors are not uniformly chosen, but with a frequency of change according to the incentives of change, then the unique stationary distribution of the continuous-time Markov chain of (local) dyad flips is described by an ERGM distribution with analogous (global) network statistics; see Snijders [2001, Section 9] for details.[15] — Note that the SAOM does not take into account the process that may have lead to the previous observed network, but is only concerned with reproducing the evolution between observations. Still, "If it is possible to reach every state from every given initial state in a finite number of steps (as is the case here), the distribution of a Markov chain with stationary intensity matrix on a finite outcome space tends to a unique limiting distribution as $t \to \infty$, independent of the initial distribution." [Snijders, 2001]

In Chapter 5 we turn to the very fundamental question to which degree tie changes in network data depend on each other. Assuming that tie changes between network observations do not conditionally depend on each other, network evolution could also be modeled with much more simple and computationally efficient statistical techniques. Within both (t)ERGM and SAOM framework we investigate, firstly, how well conditional independence and more general models explain the observed data and, secondly, whether substantive findings about network effects on the evolution of ties are affected by the (potentially invalid) assumption of conditional independence.

[15] The idea is to specify a rate function such that the denominator in Eq. 2.4 cancels away and the intensity matrix characterizing the stationary distribution reduces to $q_{ij}(y) = \exp(f_i(\theta^{(h)}, y, y^{\neg ij}))$.

2.3. Continuous Observation of Network Evolution

Network panel data, i.e. a sequence of complete network snapshots, typically contains only coarse-grained temporal information on *relational states*. A different form of temporal network data, network event data, is based on 'continuously observed' fine-grained sequences of individual time-stamped *relational events*, for which the exact minute or second of dyadic interaction between actors is known. Whereas panel data is often collected by means of longitudinal surveys, event data is typically the outcome of automated data collection such as telecommunication records. With the advent of the Internet, the availability of (computer generated) dyadic event data has grown considerably in the last years and allows for social network studies based on unprecedented large samples of the underlying population.

Crucially, as opposed to those scenarios in the previous sections, network event data is not best represented as graph(s), but rather available as a time-ordered sequence of interval-censored[16] time-stamped interaction – $(\text{time}, \text{sender}, \text{receiver})$ – tuples, (t_i, s_i, r_i), $i \in \mathbb{N}$, satisfying $t_i \leq t_{i+1}$ and $s_i \neq r_i$; the set of individuals is defined implicitly by $V = \bigcup_{i \in \mathbb{N}} \{s_i, r_i\}$. As a consequence, it is not possible to employ the large set of tools that have been developed for static and longitudinal network analysis [Wasserman and Faust, 1994, Brandes and Erlebach, 2005, Snijders, 2005] directly. If at all – 'How many (brief) events make up an (enduring) relationship?' – network ties in the sense of relational states only arise indirectly as a result of converting network event data into panel data.

At the expense of losing the ordering and spacing of original events, this conversion is usually carried out by defining a sequence of time slices and aggregating relational events within these time windows into static (weighted) networks. Arguably, however, if (at all) fine-grained event data needs to be transformed into coarse-gained panel data, then (at least) the ordering and spacing of events should feature prominently because both might contain valuable information. We touch upon corresponding models and the statistical analysis of dyadic event data in Section 2.3.2. A general review on temporal network approaches that exploit detailed information when "edges are not continuously active" has been recently given by Holme and Saramäki [2012]. In Section 6 we take another step in that direction by defining a parametric version of social vector clocks. One illustrative consequence of our approach is the increased accuracy in predicting previously unobserved interactions.

2.3.1. Example: Message Exchange on Twitter

Prototypical examples for network event data are often proprietary, such as email and phone logs. A remarkable exception in this regard is the email corpus that has been made publicly available during the legal investigations concerning the Enron corporation [Klimt and Yang, 2004].

[16] Similar to panel data, interval-censored relational event data (such as email correspondence with timestamps given by the day) does not encode the order of events that happen within the same time interval.

Here we use a different example, Twitter, since micro-blogging platforms make a great case in point for circumscribing different kind of temporal information within empirical social network data.

First, the follower relation on Twitter is similar to the friendship relation on Facebook and provides relational states of interconnected user profiles — note however that follower relations on Twitter are directed and thus allow for very different kind of analyses.

Second, the temporal dimension that is introduced through impersonally broadcasting messages to one's followers can be best understood as an additional process that takes place on the network, rather than forming it. This is *not* the kind of event data we have in mind.

Third, there is also temporal information in Twitter that has the format of dyadic event streams. That is, Twitter also supports more targeted forms of communication, in which users explicitly refer to each other. This targeted (although public) communication occurs in the form of retweets (in which one user rebroadcasts another users tweet, and attributes the tweet to its source) and user mentions, where the @ symbol is used to explicitly refer to a user. Dyadic event data – a sequence (t_i, s_i, r_i), $i \in \mathbb{N}$, of (time,sender,receiver) tuples satisfying $t_i \leq t_{i+1}$ and $s_i \neq r_i$ – can thus be extracted from Twitter log files if we

- include only this targeted form of communication (i.e., those log entries containing retweets or user mentions),
- remove self loops, and
- if a tweet mentions more than one user turn it into as many events as there are users mentioned in the tweet.

Note that, among others, this is exactly the type of data that we will use in our investigations in Chapter 6.

2.3.2. Statistical Analysis of Dyadic Event Data

The complexity of stochastic actor-oriented models for network panel data (cf. Section 2.2.3) results from the simulation of an *unknown* sequence of change events – 'micro-steps' – by means of randomly disturbed local best response decisions. For social network event data, in contrast, a sequence of dyadic interactions is explicitly *known*. In this sense, event data provide more detailed information on the network evolution than panel data.[17]

By exploiting a given event sequence, consequently, it is less complicated to account for statistical associations among future and past interactions. In particular, there is no need to establish causality; past events typically do not depend on future events and thus there is no interdependence in this respect.

[17] Note however the important distinction that dyadic observations do not encode enduring relational states but continuously observed short-lived relational events between actors.

To date, there is no predominant framework for modeling network event data (compared with ERGMs and SAOMs for analyzing one or more network observations, respectively). Various approaches exist, though. Recently proposed methods include, e.g., a multilevel discrete-time event history model for describing the occurrence and timing of relational events [de Nooy, 2011], the modeling of relational events via latent classes [DuBois and Smyth, 2010], as well as an adaptation of the SAOM framework to make it suitable for network event data [Stadtfeld, 2012] — as indicated above, parameter estimation in this setting is significantly less complex due to the given event sequence. Another line of research concerns the prediction of future events in network data; cf. Section 6.3.4.

Among the most recognized approaches is a framework proposed by Butts [2008] for modeling the *rate* of relational events. In a closely related model, Brandes et al. [2009c] also allow to analyze the *quality* of events, conditioned on (given) event occurrences. To do so, broadly speaking, each event is subsequently embedded in the *event network* of past interactions. Simultaneously, values of local network statistics which might have influenced the event are recorded. Then, once all events have been processed, statistical associations among the calculated statistics and the observed event quality (or rate, respectively) are tested; thereby, the model is based on the Markovian assumption that each event is only dependent on the directly preceding network state.[18]

Similar to the standard conversation scenario, the weighted event network in Brandes et al. [2009c], as well as the one in Stadtfeld [2012], is based on *dyadic* summations of past events (crucially, however, the impact of past events is exponentially decayed and thus reflecting the spacing of events). We note in passing that, in principle, the basic model assumptions could be combined with very different temporal network structures — which would, of course, change the interpretation of explanatory statistics, then.

2.3.3. Visualization of Dyadic Event Data

As opposed to various approaches for modeling dyadic event data, research on visualizing dyadic event data has not been developed far yet. Only recently, for instance, Binucci et al. [2012] have pointed out the underdevelopment of graph drawing algorithms that can cope with the requirements of streaming applications.

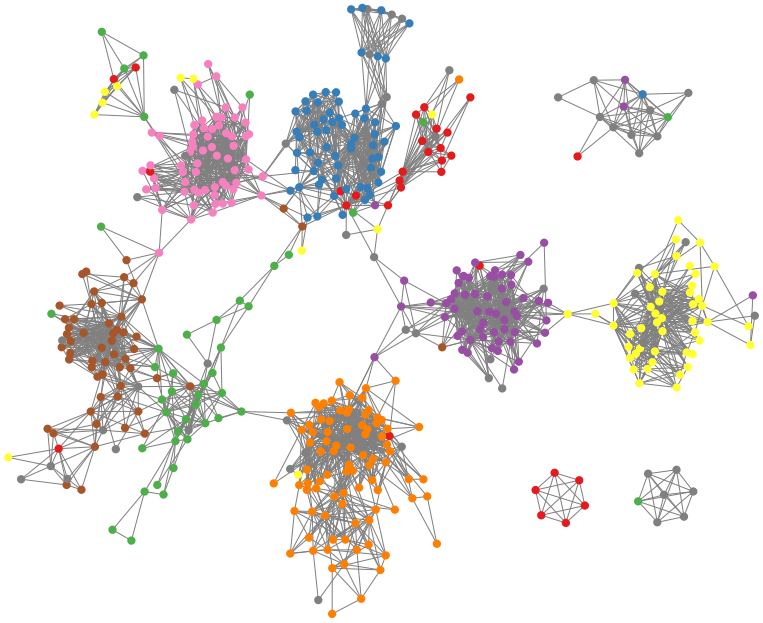
With few exceptions (such as the combination of classical vector clocks and multidimensional scaling in Harrigan [2010]) the current practice is rather to ensure applicability of time-varying network visualizations (cf. Section 2.2.2) by means of transforming a given event stream into a coarse-grained sequence of time-sliced static graphs [Bender-deMoll and McFarland, 2005] — as already noted above, we introduce the concept of social vector clocks and further elaborate on the issue of transforming event data into panel data in Chapter 6.

[18] In particular, given the event preceding network state, events in the same time unit are assumed to be conditionally independent; cf. Section 5.1.3.

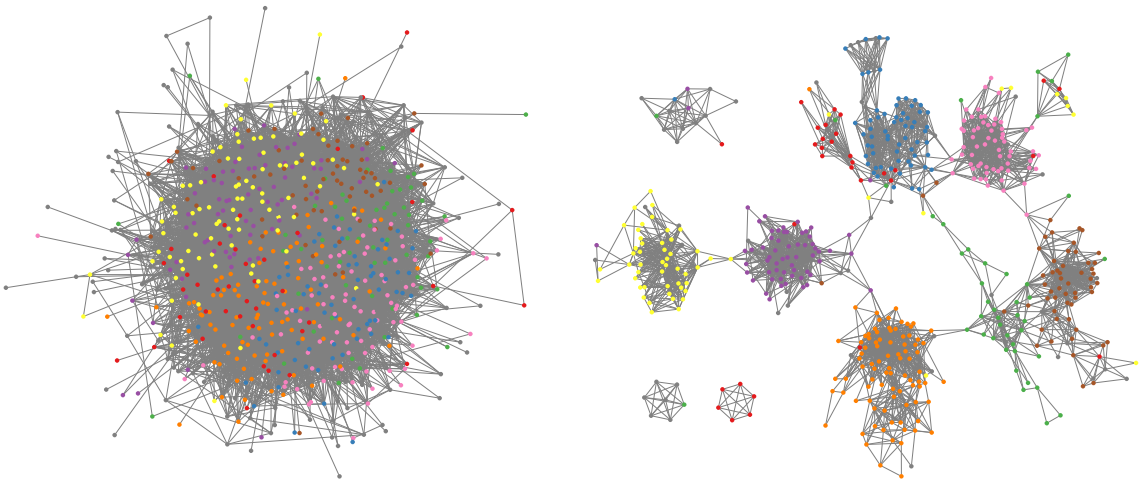
Chapter 3.

Simmelian Backbones: Identifying Essential Structure

3



The research presented in this chapter has been published in Nick et al. [2013].



(a) original network (b) Simmelian backbone

Figure 3.1.: Illustrating example. Facebook friends at California Institute of Technology, September 2005. Color encodes dormitory attribute; gray for missing value. (a) giant component as analyzed in Traud et al. [2011, 2012]. (b) reduced network according to "at least 5 overlapping top-10 ranked neighbors" w.r.t. Simmelian tie strength and after removing ties that are no longer Simmelian in the backbone.

NETWORK ties are often aggregate proxies for several heterogeneous relations. Simplifying homogeneity assumptions therefore often hamper a thorough understanding of evolving social networks — freely adapted from Orwell: all ties are equal, but some ties are more equal than others.

In online social networks, for instance, interactions related to friendship, kinship, business, interests, and other relationships may all be represented as catchall "friendships." Because several relations are mingled into one, the resulting networks exhibit relatively high and uniform density. As a consequence, the variation in positional differences and local cohesion may be too small for reliable analysis.

As opposed to this, in this chapter we introduce a novel method to identify the essential relationships in networks representing social interactions. Backed by sociological theory, we shift the network perspective from an unconditional dyadic view to a conditional triadic interpretation of cohesion, and posit that social communities are formed by *strong* and highly *redundant* — and therefore essential — ties between group members. The approach is very flexible and can be applied to directed or undirected, weighted or unweighted, and static or dynamic network data alike.

As an example, reconsider the node-link diagram of the smallest network in the `Facebook100` data (cf. Section 2.1.1), where individuals are colored according to the dorm they reside in. While the layout in Figure 3.1(a) suggests that clusters may be aligned with dorms to some extent, this property is more pronounced in the substructure shown in Figure 3.1(b). We hasten to add that this substructure is derived from the structure alone, i.e., without taking the dorm attribute into account.

In the next section, we motivate our approach by substantive theory on social groups and structural embeddedness. Computational approaches to extract essential network structure, *Simmelian backbones*, are introduced in Section 3.2, and evaluated in Section 3.3.

3.1. Sociological Theory vs. Computational Praxis

To adavance the understanding of evolving social networks, we are interested in restricting empirical social network data to those relationships most strongly embedded in social groups. While there are many conceptions of social groups, "primary" groups are of particular interest:

> Over the years sociologists have distinguished various kinds, or what Simmel called "forms" of human groups. Among these, one form in particular has continued to interest investigators for more than a century. Groups that are relatively small, informal, and involve close personal ties – those that Tönnies characterized as based on *Gemeinschaft*, Durkheim portrayed as reflecting *solidarité organique*, and both Spencer and Cooley described as *primary* – remain at the core of the discipline. [Freeman, 1992]

We focus therefore on the identification of strong ties that are embedded in primary groups. While common technical terms include, for instance, clusters and modules we henceforth refer to these primary groups as *communities*.

Note that our goal is *not* to assign community memberships themselves. As we will argue below, structural embeddedness can be defined locally, without knowing the surrounding communities or even postulating a formal definition for them. The reduction to strong, embedded ties is thus substantially different from common approaches to community detection which assess the relation between internally dense and externally sparse actor groups [Fortunato, 2010].

3.1.1. Tie Dependencies

Community detection algorithms generally equate communities with relatively dense subgroups, where 'relatively dense' means that within-group density is higher than expected with respect to some null model or substantially higher than between-group density. Modularity's null model, for example, is based on a planted partition model (cf. Section 2.1.3). More precisely, ties are assumed to form independently of each other according to the products of expected (incident) node degrees.

As we have already alluded to above, however, the processes that drive the formation of social networks go beyond the level of dyadic variables. Rather, ties are systematically patterned and embedded in local structures, such as triangles and cycles. The presence of ties is likely to conditionally depend on the presence of other ties. A typical example is the relevance of common neighbors. Indeed, such conditional dependencies fundamentally drive network evolution and "Without dependence among ties, there is no emergent network structure." [Brandes et al., 2013c]

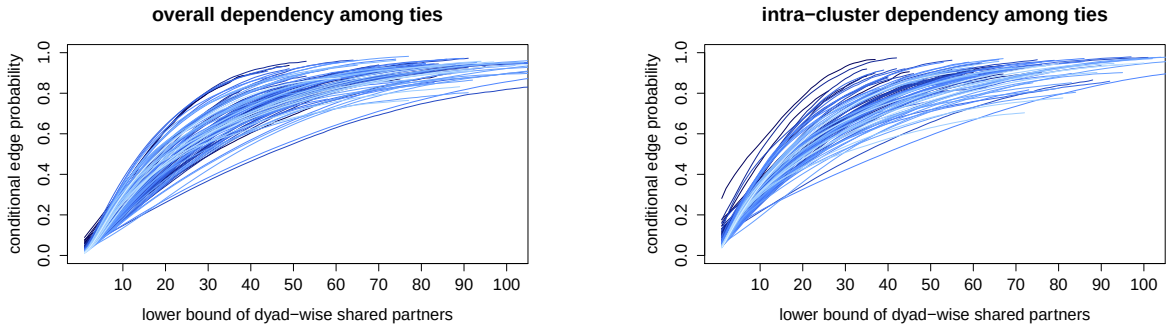


Figure 3.2.: Conditional tie probabilities in networks of Facebook friends. Each curve corresponds to one network in the `Facebook100` data (cf. Section 2.1.1). The x-axis conditions on a minimum number of triangles a dyad is contained in, and the y-axis gives the fraction of ties in such dyads. Lines are saturated according to network density which is decreasing with network size since average degree is relatively constant across networks. In the right figure, dyads are restricted to lie within communities found by Louvain modularity maximization [Blondel et al., 2008].

Since the design of valid network models obviously requires careful consideration of tie dependencies (cf. Section 2.1.3), it is rather surprising that community detection approaches such as modularity do not seek to exploit these principles; instead, state-of-the-art techniques are usually based on unconditional dyadic aggregate information, such as density and nodal degrees.

While the "cohesive" groups detected by such community detection methods do indeed correspond to relatively dense substructures, they can not be expected to exhibit other distinguished qualities that have been suggested by substantive sociological arguments, such as high transitivity.

We illustrate this point in Figure 3.2, which provides descriptive statistics on the interdependency of dyadic relationships in all networks in the `Facebook100` data. In each network, the probability of observing a tie conditioned on the number of shared friends increases with the number of shared friends. The number of shared friends equals the number of triangles a dyad can close and is sometimes referred to as its *structural embeddedness*. The effect is roughly linear for small to moderate numbers of common neighbors and then saturates.[1] When the networks are partitioned into density-based communities (using the hierarchical Louvain modularity maximization approach [Blondel et al., 2008]), conditional probabilities of intra-community dyads are lifted (which is consistent with increased density) but the functional form remains largely the same; still, for instance, a small number of common friends significantly increases conditional probabilities of intra-community ties.

[1] For this reason, the specification of transitivity effects in exponential-family random graph models commonly involves geometrically-weighted shared partner statistics; cf. Section 2.1.3.

3.1.2. Simmelian Ties and Social Groups

Local density in the sense of modularity is only one of many possible proxies for the cohesiveness of relationships among group members. We now show that sociological theory suggests that structural embeddedness (i.e., the number of triangles in which a tie is embedded, or the number of common neighbors shared by the endpoints of a tie) better accounts for the quantitative and qualitative differences between ties that exist within communities and those that exist between communities. Backed by sociological theory, we thus shift the perspective from an unconditional dyadic view to a conditional triadic interpretation of cohesion, and posit that (social) communities are formed by *strong* and highly *redundant* ties between members.

In social networks, dyadic relationships are assumed to depend on the structural environment in which they are embedded [Brandes et al., 2013c]. Network formation theories attach particular importance to triadic settings [Krackhardt, 1998, 1999]. According to Simmel [1950b], *triads* (sets of three actors) are fundamentally different from *dyads* (sets of two actors) by way of introducing mediating effects. On the other hand, "the further extension to four or more persons by no means correspondingly modifies the group any further." [Simmel, 1950a] Krackhardt goes so far to conclude that "the key to understanding the quality of a tie between two actors can be reduced to asking whether it is part of a strong triad or not" [Krackhardt, 1998] and classifies dyadic relationships accordingly: "super-strong" *Simmelian ties* are embedded in at least one triangle of reciprocated ties, while sole-symmetric and asymmetric ties are not; cf. Figure 3.3.

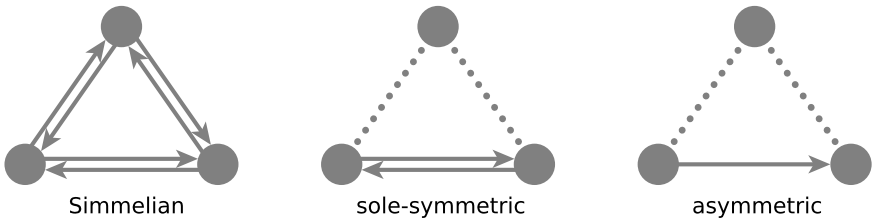


Figure 3.3.: Qualitatively distinct triadic embeddedness of dyadic relations according to Krackhardt [1998]: "super-strong" *Simmelian ties* are embedded in at least one triangle of reciprocated ties, while sole-symmetric and asymmetric ties are not.

Prominent examples of triadic perspectives include Heider's structural balance theory [Heider, 1946, Krackhardt and Handcock, 2007] and Granovetter's theory of the strength of weak ties [Granovetter, 1973]. As others have noted, "Granovetter's argument contends that weak ties provide better connections to different social milieus because they usually connect socially dissimilar people" [Wellman and Gulia, 1999].

Similar to Granovetter, Burt distinguishes between weak ties that provide non-redundant connections to information buried in structural holes, and strong ties, which are much more interconnected (i.e., highly redundant) [Burt, 1992]. A group's capabilities (and constraints) are then determined by two dimensions: *the redundancy of internal ties* (along the horizontal axis of Figure 3.4) and *the non-redundancy of ex-*

ternal ties (along the vertical axis of Figure 3.4). Burt argues that a group achieves maximum performance when its internal ties are strong, and its external ties are weak: "while brokerage across structural holes is the source of added value, closure can be critical to realizing the value buried in the structural holes" [Burt, 2001]; cf. Figure 3.4.

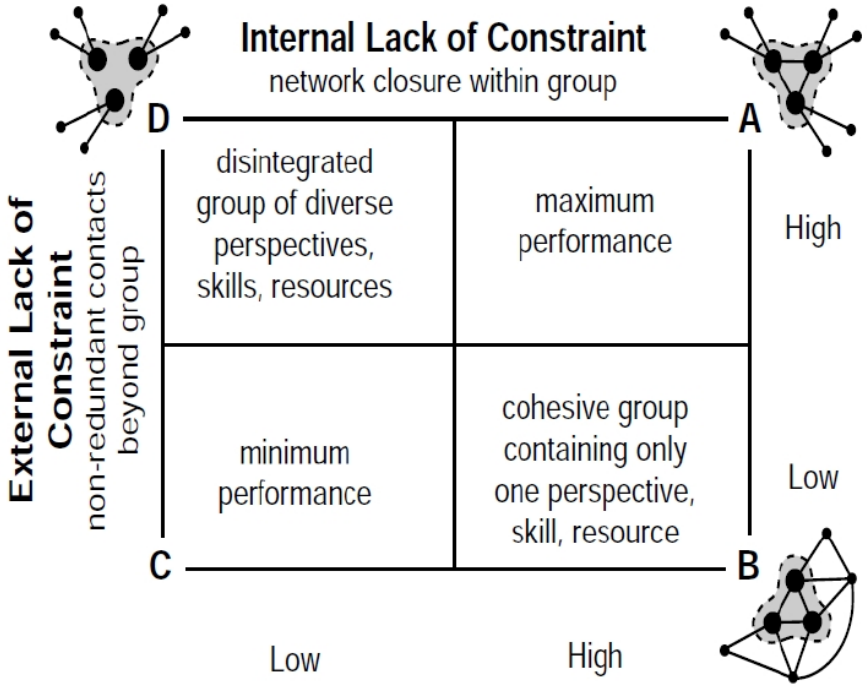


Figure 3.4.: Group performance according to Burt's notion of constraint: "while brokerage across structural holes is the source of added value, closure can be critical to realizing the value buried in the structural holes" (reproduced from Burt [2001]).

Our intended notion of triadic cohesion is driven by this conception of a high-performing group with strong internal connections and weak external connections, as in the upper-right quadrant of Figure 3.4. This notion, which is fundamentally based on the embeddedness of ties, is thus very different from aggregate density-based measures such as modularity, which do not consider embeddedness at all.

Against this background, we now specify a method for distinguishing

a) intra-cluster ties: "strong", Simmelian, structurally embedded ties exhibiting a high level of redundancy and homogeneity, from

b) inter-cluster ties: "weak", non-Simmelian, less embedded ties that are rather diverse and heterogeneous.

3.2. Extracting Simmelian Backbones

In the following we propose a framework to extract a substructure consisting of ties that are strong and redundant. The approach is based on a triadic notion of cohesion and consists of the following steps:

step 1. If input tie strength is uniform, assign to ties the number of triangles they are embedded in (Simmelian strength).

step 2. For each actor, rank all alters by tie strength.

step 3. For each (strong) tie, determine its redundancy.

step 4. Filter ties that are weak or not redundant.

3.2.1. Tie Strength

Our procedure requires that each tie is labeled with a preliminary strength that is a proxy for social importance. While such weights are rarely explicitly known, common proxies are frequency of contact in communication networks or shared attributes in recommender systems. If such a tie strength is available, then one can simply use this and proceed to the next step.

Even if there is no explicit information on tie strength, as is the case with Facebook friendship relations, sociologically informed weights can be extracted from the network structure itself. According to Granovetter, structural embeddedness is an important proxy for tie strength and in many substantive applications embeddedness can be characterized by triadic configurations.

While the specific weighting technique will depend on what ties represent, a generic proxy for uniformly weighted friendship networks is Simmelianness [Krackhardt, 1998]. Dekker [2006] lists several tie-level measures of Simmelianness and claims that "The main advantage of the Simmelian tie strength measures is that they are firmly rooted in a substantive theory." Under the binary variant, a tie is considered Simmelian if it is part of a triangle (cf. Figure 3.3). A straightforward extension to measure the degree of Simmelianness for a tie $\{i, j\}$ is to count the number of neighbors shared by i and j (i.e., the number of distinct triangles a tie is embedded in). Dekker argues that shared partners "are central in opinion formation about behavior and intentions" and ever more shared partner impute constraints "on each others behavior, because more of their behaviors become observable to each other." [Dekker, 2006]

For the purpose of illustration, and the examples in the next section, we use this weighting technique (i.e., the number of shared neighbors) as a preliminary proxy for social importance in unweighted social networks. We discuss the computational complexity of this approach and the handling of other types of network data below.

3.2.2. Ranking Ties

A pivotal factor in our approach is the following: In correspondence with Burt's notion of constraint, we do not consider a global network-level ranking of ties but advocate a local assessment on the level of actor neighborhoods. Let ego be an actor in a network with ordinal tie weights. It does not matter whether the relation is directed or not. We define ego's *rank-ordered neighborhood* as the list of adjacent actors (alters) sorted from strongest to weakest tie between ego and alter.

$N^{\rightarrow}_{ego}$: ego's alters ranked from strongly to weakly tied

This actor-level ranking procedure is very different from ranking all ties globally, for at least two reasons. First, the procedure introduces *asymmetry* between connected pairs of nodes, even in undirected networks: for example, whereas B may only rank seventh in A's list, A may rank first in B's list. Secondly, the procedure can deal with *heterogeneity*: ties which have the same strength from a global perspective may have very different rankings from a local perspective. For example, a tie embedded in ten triangles in a very dense region of the graph may rank low, whereas an equally embedded tie in a sparser portion of the graph may rank high. Whether the input was directed or not, the result of this step is a (rank-weighted) directed graph which we refer to as the *ranked neighborhood graph*.

3.2.3. Tie Redundancy (Structural Embeddedness)

There are various ways of filtering the ties of a (weighted) graph based on derived weights and absolute or relative thresholding (see, e.g., Goldberg and Roth [2003], Radicchi et al. [2004], and Melançon and Sallaberry [2008]). Many of these, however, bear little relationship with the Simmelian idea of groups.

Based on the ranked neighborhood graph, we therefore introduce a novel measure of triadic cohesion. The crucial idea is to compare the local perspectives of ego and alter on the social importance of others: a strong (i.e., highly ranked) tie is considered strongly embedded,[2] if its endpoints have similar views on highly ranked neighbors — that is, the actors incident to a *triadic cohesive tie* are relatively strongly connected to each other and also relatively strongly connected to relatively many common neighbors (and thus embedded in very strong Simmelian groups).

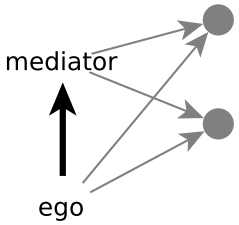
Formal definitions of the degree of embeddedness thereby correspond to a similarity assessment of ranked neighborhood lists. In this way, we can avoid that (additional) weak ties to distinct neighbors degrade the evaluation of top-ranked (strong) redundancies. Two examples, one parametric and one non-parametric, follow below.

[2] In the present setting we find the term (structural) embeddedness more intuitive than redundancy.

Parametric variant

A granularity parameter k specifies the rank after which ties are not considered to be relatively strong anymore. The degree to which a strong tie is structurally embedded is then defined as the *overlap of common neighbors among the top-k entries* of $N^{\rightarrow}_{ego}$ and $N^{\rightarrow}_{alter}$.

This procedure corresponds to evaluating ego's (strong) tied alters in the directed top-k ranked neighborhood graph w.r.t. mediating positions in transitive triplets; see adjacent diagram. Transitive (mediated) triplets and balanced neighborhoods have proven crucial, e.g., in estimating ERGMs and SAOMs (cf. Chapter 2).



A general reference for the comparison of top-k lists is Fagin et al. [2003].

Non-parametric variant

While a granularity parameter k appears convenient for exploring structural embeddedness at different levels, it may also prove to be a burden in the absence of substantive expectations on the degree of embeddedness.

A non-parametric alternative is to automatically determine, for each tie, a k that maximizes the overlap among the top-k entries of $N^{\rightarrow}_{ego}$ and $N^{\rightarrow}_{alter}$ as above. Because this involves prefixes of different length, however, the number of common entries is normalized by the size of the union of entries. In other words, we compute the *Jaccard coefficient for prefixes of any length* and pick the maximum value. Note that the classical Jaccard coefficient (i.e. the comparison of all neighbors) only provides a lower bound for the best prefix Jaccard coefficient.

While the non-parametric variant allows for even more heterogeneity across ties, we do not yet have a good grasp of the consequences because of its more difficult interpretation.

We note in passing that both variants can take into account reciprocity in the redundancy assessment by identifying alter's rank in $N^{\rightarrow}_{ego}$ with ego's rank in $N^{\rightarrow}_{alter}$.

3.2.4. Filtering Ties

We define the *Simmelian backbone* of a social network by applying a (global) threshold on the derived (local) measure of triadic cohesion. Minimum requirements include, e.g.,

- a minimum number of common top-k neighbors, or
- a required prefix Jaccard coefficient.

In other words, we filter the ties of a ranked neighborhood graph first by their relative strength (i.e., their rank in $N^{\rightarrow}_{ego}$) and then, additionally, by their strong embeddedness (i.e., the similarity of $N^{\rightarrow}_{ego}$ and $N^{\rightarrow}_{alter}$ in their top ranks).

In the so reduced network, remaining ties are directly and indirectly attached to top-ranked neighbors: we expect to observe high similarity between the ties of ego and the ties of those alters to whom ego still has a tie after the filtering (while weak and less embedded connections are implicitly filtered away). Based on this transformation, therefore, we induce an individual-level attachment to groups corresponding to Burt's notion of constraint. Indeed, the proposed heuristic can be interpreted as a pairwise comparison of ego networks, since all calculations are solely based on local information.

As we will demonstrate in Section 3.3, social communities and bridges between them become rather obvious in the filtered network. For convenience, inter-community bridges can be filtered by removing also ties that are no longer Simmelian in the backbone. However, where appropriate, this should not be done for ties in the 1-core that are not in the 2-core, since these connect otherwise isolate actors with the community to which they are most attached.

If the communities themselves are of interest, any technique from the broad range of existing community detection algorithms can be applied to partition the network via its backbone. As illustrated in the next section, the balance between inter- and intra-community ties has altered strongly in favor of the latter, so that community detection is easy.

We conclude this section with a word on the computational complexity of this approach, including the weighting scheme. The triangles of a graph G can be listed in time $\mathcal{O}(m\alpha(G))$ [Chiba and Nishizeki, 1985], where m is the number of ties and $\alpha(G) \leq \sqrt{m}$ its arboricity. The arboricity is defined as the minimum number of forests needed to cover the ties. Since it is bounded from above by the h-index of the graph and the latter has been found to be small for networks of social relations [Eppstein and Spiro, 2009], the arboricity can be expected to be small as well. Both Simmelian tie strength and the degree of embeddedness can be determined by a slight variation of the triangle listing algorithm. In the non-parametric variant the latter requires $\mathcal{O}(m\alpha(G))$, again, and in the parametric one it becomes linear for fixed k. Moreover, neighborhoods can be rank-ordered in linear time using bucket sort, since tie weights are bounded by $n - 2$, the maximum number of shared neighbors in a graph with n nodes. Overall, the ranked neighborhood graph of an unweighted graph (including the weighting technique described above) and its Simmelian backbone is constructed in $\mathcal{O}(m\alpha(G))$ time where $\alpha(G)$ is expected to be small.

3.3. Case Study: Amplifying Hidden Homophily in Facebook Networks

We illustrate a potential use case of Simmelian backbones by re-analyzing the 100 networks of Facebook friendships in the `Facebook100` dataset; cf. Section 2.1.1. Traud et al. [2011] used a modularity maximization technique to detect communities in their investigation of the `Facebook100` networks, and found that in some cases dorms

displayed a high level of homophily within communities, and in other cases the year of graduation appeared to be the attribute most related to community structure. In particular, the dorm attribute tended to have high homophily in communities found at small institutions, whereas the year attribute was more enriched in the communities detected at larger institutions. Findings in Lee and Cunningham [2013] indicate that this tendency may be an artifact caused by limitations of modularity-based community detection, which is known to suffer from a resolution limit [Fortunato and Barthelemy, 2007].[3] Thus, despite previous work on the topic, the relationship between the community structure, and the level of homophily for each of the node attributes, remains somewhat uncertain.

3.3.1. Visual Exploration

Visualization of these networks has proven difficult. Three of the `Facebook100` datasets were used for the Visual Analysis of Complex Networks (VACN) challenge.[4] While some interesting visualizations emerged, none of them clearly displayed both the connections between and within communities.

We provide anecdotal evidence that extracting a Simmelian backbone and visualizing the resulting subgraph yields more informative visualizations of the community structure present in `Facebook100` networks. In addition to the Simmelian backbone of the California Institute of Technology (Caltech) network visualized in Figure 3.1(b), we show the backbone of the University of Chicago network in Figure 3.5.

The Caltech network received extensive discussion in Traud et al. [2011], in which the authors note that at Caltech the residential house system is particularly important for the formation of social relationships. However, the network visualization of the Caltech network presented in that paper [Traud et al., 2011, Figure 2.1] does not provide a clear picture of how the dormitories are closely related to social life; this relationship is much clearer in Figure 3.1(b).

In the following, we will pay particular attention to the Simmelian backbone of the University of Chicago network,[5] which includes 6,591 nodes and 208,103 undirected ties. The residential "houses" are known to be of utmost importance to the social life at the University of Chicago: upon entering the university, every student is required to spend at least one year living in the dormitory system, and the friendships formed in this stage of college—for many students, the first time living away from home—often endure for years. According to the University of Chicago website, "each house represents a tight-knit community of students, resident faculty masters, and residential

[3] For example, maximizing modularity in many cases produced unreasonably large and few communities. A parameterized version of modularity was shown to alleviate these problems, but Traud et al. [2011] did not use the parameterized definition in their investigation.

[4] With participation from tools Gephi, i2's Analyst Notebook, Pajek, Tulip, and visone. See `http://sociograph.blogspot.com/2011/02/visualizing-large-facebook-friendship.html` for details including the two best visualizations of the California Institute of Technology (Caltech) network – compare these to Figure 3.1(b).

[5] One of the co-authors in Nick et al. [2013] has first-hand experience of the social life there and, ironically, his Facebook profile is included in the analyzed Chicago network.

(a) Nodes colored by dorm

(b) Nodes colored by year of graduation

Figure 3.5.: Giant component of Simmelian backbone of Facebook friends at the University of Chicago (3854 nodes, 19724 edges): (a) Nodes colored according to dorm they live in. Because there are dozens of distinct values for this attribute, not all colors are visibly distinct. (b) Nodes colored according to year of graduation. Note the new arrivals on the right (red colored). Grey nodes indicate missing attribute values.

staff, who live, relax, study, dine together at House Tables, engage, socialize, and learn from each other." [6] [7]

As we can see by zooming in closely on Figure 3.5(a), in which nodes are colored by their house, the Simmelian backbone does a good job of recognizing clustered structures in which the dorm attribute displays a high level of homophily, providing a strong validation of the approach. The major exception to this trend is the large cluster located at the 3:30 position. In Figure 3.5(b) we see that all of these nodes share the same year; in fact, these red colored nodes have 2009 as their graduation year. It happens to be the case that the `Facebook100` dataset was collected in September 2005. In the middle of that month, the University of Chicago class of 2009 moved into their dorms and attended orientation week, during which they interacted with each other and had very little interaction with students graduating in other years, most of whom would not arrive back in Chicago until the end of September when the Fall term started. It is interesting to see the first-year students isolated in their own group, having not yet met and integrated with the older members of their dorms. We note that this feature is not apparent in visualizations which include all ties in the graph—we first noticed it when looking at the Simmelian backbone.

3.3.2. Quantitative Description

We persistently observe the same effect for almost all network and all relevant attributes, namely that backbone ties display a stronger tendency for homophily than the entire set of ties. The share of homophily in the original network and in the Simmelian backbone for each `Facebook100` network is contrasted in Figure 3.6.

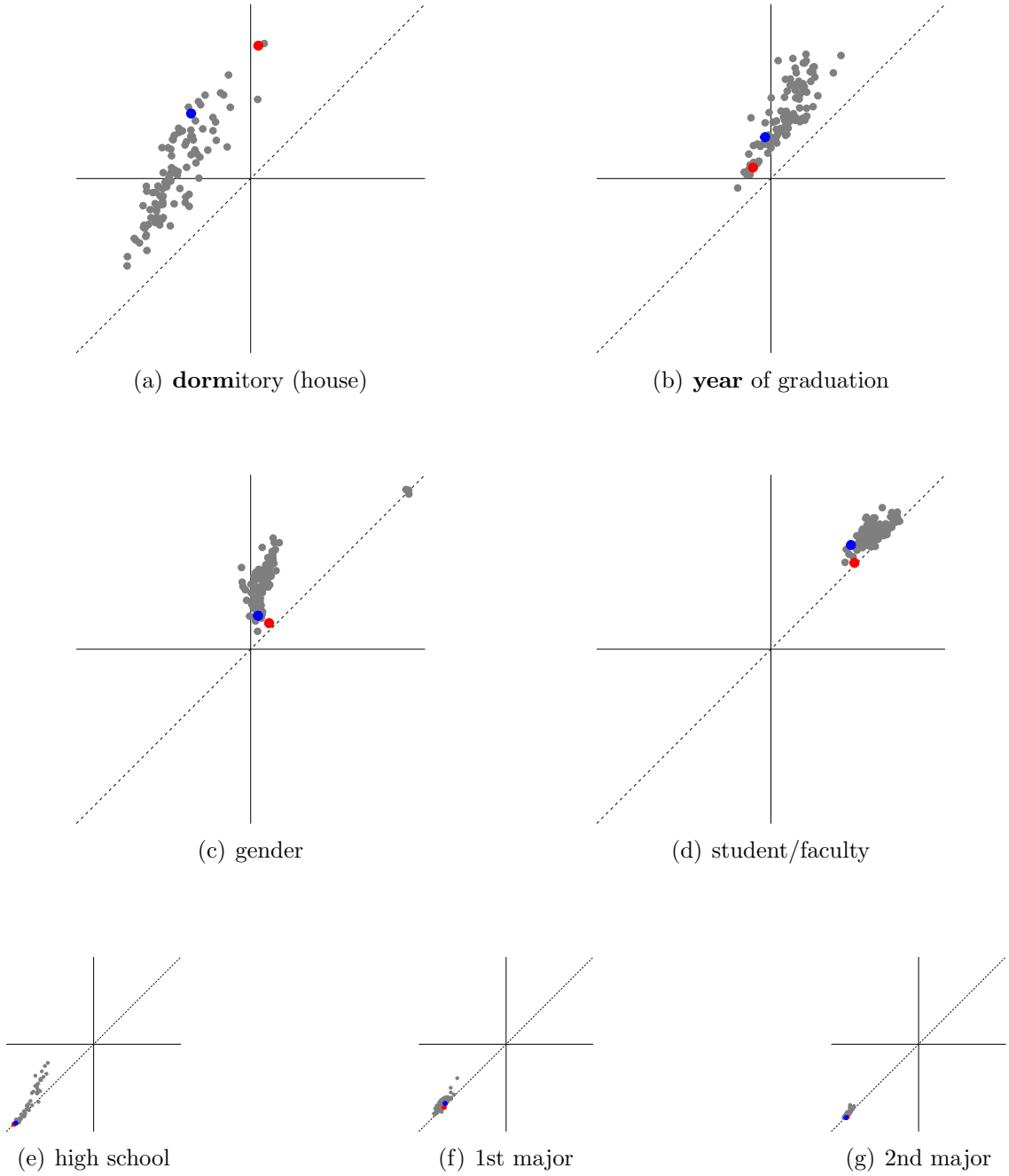
Each point in the scatterplots represents one institution in the `Facebook100` data. Points above the dotted line through the origin indicate that homophily in the Simmelian backbone is increased compared to the original network with respect to a focal attribute. In detail, for each attribute, ties are classified as follows: "homophily ties" connect students having the same attribute value, while "heterophily ties" connect students having a different attribute value. Then, the share of homophily

$$\frac{\#\text{homophily ties}}{\#\text{homophily ties} + \#\text{heterophily ties}},$$

in the original network (x-axis) is contrasted with the same fraction of ties in the Simmelian backbone (y-axis); centered at $(0.5, 0.5)$. Ties with attached student for which the focal attribute is undefined are not taken into account in this ratio calculation.

[6] `http://csl.uchicago.edu/feature/announcing-new-residence-hall-and-dining-commons`

[7] The meta-data does not explicitly indicate that the "dorm" attribute represents the housing system at the University of Chicago, but this is strongly suggested by the data itself, as the number of distinct values corresponds quite closely to the number of houses and fraternity houses. At the time of data collection, the housing system consisted of 10 residence halls (physical buildings) which were further subdivided into 37 houses, which typically represent a physically adjacent wing of a residence hall. House sizes range from 100 to 37 students, with an average size of 70. Furthermore, there were somewhere between ten and twenty fraternity houses. Thus some of the dorm information may indicate fraternity membership.



(a) **dorm**itory (house)

(b) **year** of graduation

(c) gender

(d) student/faculty

(e) high school

(f) 1st major

(g) 2nd major

Figure 3.6.: Share of ties linking nodes with the same attribute value in original network (x-axis) and Simmelian backbone (y-axis). Each dot represents one network of Facebook friends with Caltech (red) and University of Chicago (blue) highlighted. For networks above the diagonal, homophily is stronger in the backbone.

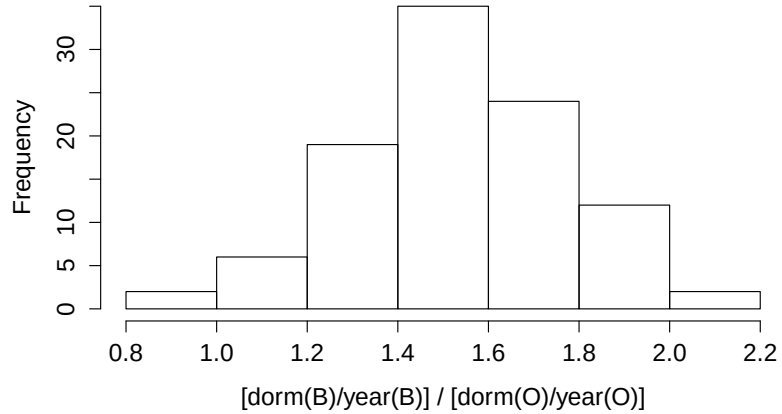


Figure 3.7.: The prominence of dorm-homophily over year-homophily increases on average by a factor of 1.5 in the Simmelian backbones (B) when compared to the original networks (O).

The scatterplots in Figure 3.6 indicate which attributes are most important factors in network formation on the aggregate level: dormitory, graduation year, gender, and student/faculty status. Our Simmelian backbones reveal that homophily in the first three attributes is much stronger than can be noticed from the original networks. In addition, they also show that the relative importance of attributes is shifted after denoising the networks. Figure 3.7 in particular suggests that dorm-homophily, which could previously only be observed in the smaller networks [Traud et al., 2011, 2012], was subdued by year-homophily.

These numbers together with the above network visualizations clearly suggest that within-dormitory communities are more prominent than could previously be detected.

3.4. Discussion

We presented an approach to uncover deeply embedded relations that are likely to be internal to primary groups, referred to as communities. The method serves, among other things, to reveal essential structure in networks with insufficient variation in local density for standard community detection approaches to work.

The main ingredient of our method is the concept of *ranked* neighborhood comparisons to infer the structural embeddedness of ties. The corresponding notion of triadic cohesion has three desirable aspects: (i) it implements sociological theory and should therefore be easier to justify in empirical settings, (ii) it is defined locally and therefore applicable to ego networks as well, and (iii) it is computationally efficient and scales to big network data.

As a prototypical application, we showed that Simmelian backbones of Facebook friendship networks may serve to reveal primary groups that provided strong evidence of homophily and had been difficult to detect previously.

For the clarity of exposition we have restricted the illustration of Simmelian back-

bones to static unweighted undirected network data. However, the method can be adapted straightforwardly to directed, weighted, and temporal network data as well — only the extraction of tie strength proxies in `step 1` (i.e., the utilization of the number of common neighbors) has to be altered, as appropriate.

In the next section, for instance, we will deal with *weighted* longitudinal networks that obviate the need for `step 1` in the provided framework, a priori. In particular, the Newcomb fraternity data (cf. Section 2.2.1) provide complete sociometric preference rankings which even obviate `step 2`.

As a second example, dyadic event data and aggregate networks extracted thereof (cf. Section 2.3) are typically *directed* and thus allow to distinguish Simmelian triads from other triangles (such as induced directed 3-cycles) in `step 1`. Alternatively, the number of events counted for each tie can be used as (utterly dyadic) proxy for social importance. In contrast, in Chapter 6 we provide a more elaborate algorithm to transform dyadic event data into networks with edge weights that align well with the here presented theory.

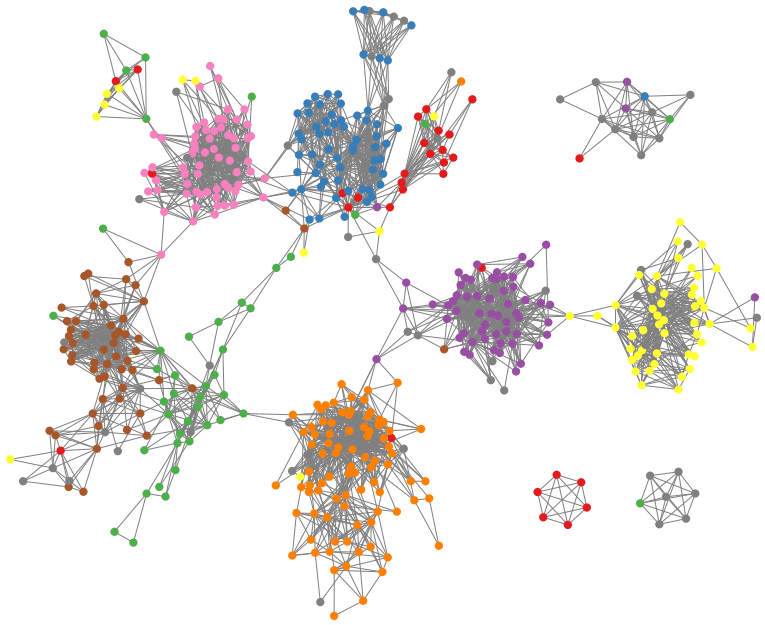
The comparison of different tie strength proxies for different kind of network data is an obvious direction for future research on Simmelian backbones. Other interesting directions include the exploration of alternative definitions of relatively strong structural embeddedness and different filtering schemes. More importantly, we are ultimately interested in structural properties of Simmelian backbones related to invariants of the input graph as well as their impact on community detection algorithms.

A free network analysis tool offering Simmelian backbone extraction is `visone`.
See `www.visone.info` and its description in the visone wiki.

Chapter 4.

Network Exploration with Gestaltlines

4



This chapter is based on Brandes and Nick [2011] and Brandes et al. [2013b], and draws upon collaborations within *argo*, the interdisciplinary transfer center for the collection, analysis, visualization, and modeling of network data at the University of Konstanz.

TOWARD a better understanding of evolving social networks, visualization is an essential tool in both the exploration of data and the communication of findings [Freeman, 2000, Brandes et al., 2006, Correa and Ma, 2011].

While cross-sectional data pose many challenges for network visualization already, longitudinal data increase the level of complexity significantly [Moody et al., 2005]. Because of this complexity and the multitude of interests that analysts may have, it is unlikely that there are general visualization schemes serving the majority of needs. Instead, social network visualizations should be tailored to the specific type of data and analytical interest at hand.

Here we are interested in such a specific scenario, which is rather typical for empirical studies of longitudinal social networks (cf. Chapter 2.2). We assume a time-dependent network of *asymmetric* relations and an interest in the evolution of dyadic (i.e., pairwise) relations along with their embedding in the structure at large.

Others have noted that a "large number of asymmetric ties suggests that we might gain some insight by using a layout method that accounts for this asymmetry" [Moody et al., 2005, page 1228]. Given highly asymmetric relational data, however, directed (node-)link representations become less comprehensible already in the static case [Holten and van Wijk, 2009].

In contrast, we aim for an intuitive way to communicate *entire dyadic evolution* clearly and effectively through simple graphical means. Patterns of interest, such as intervals of relative stability, periodicity, trends, transitions, as well as outliers, are often difficult to describe quantitatively, even in hindsight, but, as we will demonstrate, do possess relatively simple and coherent visual expressions. It goes without saying that detecting such patterns, as well as shared, related, and in particular discriminating features of exceptional actors and dyads are of great importance in identifying factors that govern the network evolution; not least because current models rely on fairly strong homogeneity assumptions (cf. Section 2.2.3).

The here proposed network visualization is a prototypical application of what we termed *gestaltlines*. It is, hence, inspired by a combination of three powerful concepts from information visualization: Tufte's sparklines, gestalt theory, and multivariate glyphs. We provide detailed background on gestaltlines in Section 4.1.

In Section 4.2, then, we turn to the special design of gestaltlines for longitudinal network data. Initially, we illustrate the derivation of corresponding representations for asymmetric relations. Subsequently, these diagrams will be incorporated into a matrix view, to convey the entirety of dyadic time-series in network panel data in a single diagram that allows for visual exploration of patterns, trends, and outliers.

4.1. Gestaltlines

Our general motivation is the use of visualization to facilitate exploration and communication of patterns in multivariate data sequences. Patterns of interest typically include intervals of relative stability, periodicity, trends, transitions, outliers, and, in the case of sequence collections, shared, related, and discriminating features. These

are often difficult to describe quantitatively, even in hindsight, but do possess relatively simple and coherent visual expressions.

A more particular motivation is the intended display of such visualizations on high-resolution, but small scale, media — such as paper. These allow for complex designs while at the same time facilitating scrutiny within the span of the eye and data narrations within the flow of text.

4.1.1. The Concept

In our *gestaltline* approach to multivariate data sequence visualization, we advocate the conscious blending of three established concepts that explicitly leverage human pattern recognition capabilities in small space. Our approach is based on the arrangement of multivariate glyphs (see, e.g., Ward [2002] and Ware [2004, Chapter 5]) in sparklines [Tufte, 2006] to evoke gestalts (see, e.g., Sternberg [2008, Chapter 3]) that correspond to patterns in the data. We review briefly the three constituting elements before proposing a scheme to integrate them.

Sparklines

Sparklines are "data-intense, design-simple, word-sized graphics" [Tufte, 2006]. The main rationale for these *datawords* is to allow exploratory visual comparison of large amounts of data *within eyespan*. While high resolution is a prerequisite, however, display media need not be large, and often should not. The concept is best thought of as geared toward fine printing on paper.

Because of their size, sparklines have been referred to as the "Tom Thumb of Statistical Graphs" [Yokum, 2009]. Small size facilitates the arrangement of multiple aligned statistical graphics as well as their use directly within text. It allows, e.g., to show data elements repeatedly and exactly where they are referred to, thus eliminating the need to go back and forth between a figure and associated statements. Common examples include line , bar and win-loss charts of, say, measurement time series, sports results or, most prominently, stock quotes. Although we cannot quantify this claim, it appears that most applications of sparklines today involve univariate time series data.

The design space for sparklines is almost as huge as it is for any statistical chart. Given their use in small multiples and in chosen locations inside of sentences, however, slightly different goals may be pursued. For example, a tabular arrangement may call for alignment. Moreover, sparklines can be annotated to convey simple statistics such as normal ranges, as well as specific data points of special interest  glucose 6.6 [Tufte, 2006, p. 47] which may be different each time the data are referred to in a text passage.

Note, however, that area considerations may also introduce additional constraints. If a line chart is integrated into text, for instance, its height is constrained by font size; if in addition an average slope of 45 degrees [Cleveland, 1993, Heer and Agrawala,

2006, Tufte, 2006] is desired, then the length of the corresponding dataword is implied as well.

Glyphs

The term *glyph* is used to refer to a class of graphical objects with several degrees of freedom that can be used to represent multidimensional data points by mapping each dimension of the data to a distinct free parameter. Comprehensive introductions and general design guidelines can be found in Ward [2002], Ware [2004].

The main rationale is that uniform depictions of multiple attribute values in a single, complex graphical object are easier to memorize and compare than groups of simpler graphical objects that represent data dimensions separately. A well-known example of this kind are star plots [Chambers et al., 1983], in which each data dimension is represented along a radial line segment out of a common origin ✳. Bounding the asymmetric stars by filled polygons ◤ yields integrated, yet characteristic shapes for each multidimensional data point.

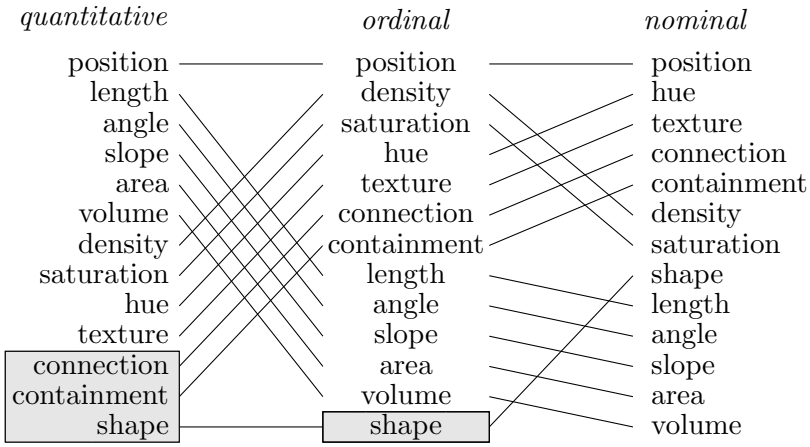


Figure 4.1.: Ranking of perception accuracy (top-to-bottom) as a guideline for graphical mapping. Boxes indicate variables irrelevant for the corresponding type of data. Redrawn from Mackinlay [1986].

Heterogeneous dimensions and varying measurement scales call for more elaborate designs. Although this is yet another sprawling topic, there are some principled guidelines. As indicated by experimental evidence summarized in Figure 4.1, the various graphical variables yield differential accuracy in elementary perceptual tasks [Cleveland and McGill, 1984, Mackinlay, 1986, Heer and Bostock, 2010]. The features of a glyph should therefore be chosen accordingly. These choices may, however, interfere when elements are perceived holistically as in the width and height of a rectangle [Ware, 2004].

Gestalt Theory

Wertheimer postulates that the mind organizes disparate visual stimuli into the simplest stable and coherent form [Wertheimer, 1923]. In other words, we are biased toward perceiving wholes, or *gestalts*, rather than collections of individual parts. Based on this so-called *Law of Prägnanz*, gestalt theory consists of qualitative principles such as those listed in Figure 4.2. A more detailed overview is given in Sternberg [2008].

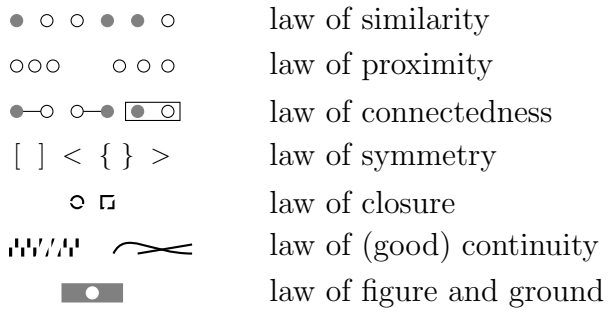


Figure 4.2.: Seven basic gestalt laws of perception. For example, the law of continuity suggests that we tend to perceive two crossing rather than two touching lines.

Applications of gestalt theory in visualization design include human-computer interaction [Flieder and Mödritscher, 2006], visual screen design [Chang et al., 2002], algorithmic animations [Esponda-Argüero, 2010], and information dashboard design [Few, 2006], as well as animated visualizations of network data [Nesbitt and Friedrich, 2002].

Gestaltlines

While the alignment of glyphs in small multiples is a natural extension to depict sequentially or spatially ordered multivariate data (see, e.g., Healey et al. [1995], Chuah and Eick [1997], Horn et al. [2001], Kosara and Miksch [2001], Fanea et al. [2005]), we posit that two additional design considerations, compactness and gestalt, will bring to bear the real potential of such visualizations. We refer to designs that result from this principle – the arrangement of gestalt theory-informed glyphs in sparklines – as *gestaltlines*.

As a restricted and univariate, but nevertheless inspirational example consider the following illustration of the workings of sorting algorithms [Sedgewick, 1998]. Each element of a data array is represented by a line segment, the slope of which corresponds to the position of that element in sorted order. An unsorted initial array is thus easily distinguished from a partially sorted intermediate array and the fully sorted final array . Note the use of slight unimodal length difference to emphasize the visual effect. By visual comparison of intermediate states, the operations of different sorting algorithms are illustrated very graphically.

The main challenge in creating gestaltlines is to identify glyphs and alignment rules from which the presence and specific nature (e.g., location, extent) as well as the absence of certain patterns can be perceived holistically.

Generally important data patterns include clustering and outliers, whereas patterns that are particularly relevant for sequence data include trends, periodicity, disruptions, change points, or phase shifts. Note that in multivariate data such patterns may emerge from combinations of dimensions.

Establishing broadly applicable guidelines for gestaltline design is going to require a major research effort because of the interdependence of glyph design, arrangements, possibly emerging gestalts and patterns of interest. Note that even such fundamental knowledge as the ranking of graphical variables in Figure 4.1 may be invalidated when we try to make certain patterns graphic by aiming at specific gestalts. In the following section, therefore, we provide selected examples of well-known data sets from the literature to illustrate some possible design choices.

4.1.2. Examples from the Literature

Before turning to the design of gestaltlines for longitudinal network data, in this section we will discuss four short examples from the non-network literature to illustrate the general potential of explicit consideration of gestalt principles in the visualization of multivariate data sequences. That is, the purpose of this discussion is to pinpoint aspects for consideration, rather than the design of the most appropriate diagrams. The examples are meant to demonstrate that using glyphs in sparklines may be only a slight generalization, if any, but that explicit consideration of gestalt principles does make a difference. The interested reader is also referred to Appendix A, in which we provide a more elaborate case study and empirical evidence for the usefulness and validity of there presented specific gestaltlines.

Phase Shifts in Population Dynamics

The dynamics of predator-prey populations are examples of bivariate data sequences. We here use a classic data set in which fur trade records indicate the population size of Canadian lynx and snowshoe hares between 1900 and 1921 (see, e.g., Odum [1971]).

Such pairs of sequence data can be depicted straightforwardly in sparklines using superimposed line  or bar  charts. The data are represented with high accuracy and comparison of the size of the two populations in any given year is easy. The dominant and well-known pattern in this data is one of periodic peaking of both populations, with predators lagging behind prey.

The same pattern is also visible in an alternative design  using pairs of dots with areas proportional to population sizes. Here, the law of similarity suggests that diagonal grouping of large dots is more immediate than vertical grouping to time. The slope of perceived diagonals is an indicator of the lag between population surges. Since the perceived diagonals are approximately parallel, the lag is roughly the same between both pairs of peaks and phase changes stick out.

So far this re-iterates what is also obvious from the line or bar chart representations, and we even paid a price because relative areas are perceived less accurately than relative positions or lengths (recall the ranking in Figure 4.1). In a longer sequence involving more peaks, however, the law of similarity also applies on another level. If there was a period in which the lag differs, this is more easily recognized as an outlier among the otherwise similar diagonals. In the extreme case that the sign of the slope is reversed  , the outlier is detected pre-attentively. A population of predators surging before the prey would be a very interesting data pattern. Due to the strong separation of dimensions when showing color coded populations in line  or bar  charts, such a pattern is more easily overlooked (compared to the case of an emergent atomic feature, as is the slope of a perceived diagonal).

Streaks in Sports Results

Tufte uses the win-loss charts for baseball teams to demonstrate that "Sparklines can simultaneously accommodate several variables" [Tufte, 2006, p. 55]. These charts contain a tick for each game  and the tick's location above or below an imaginary center line indicates whether the game was won or lost. In addition, the centerline is actually drawn when the game was played at home.

By the law of proximity, streaks of wins and losses are grouped, and by the law of connectedness, partially contradicting stretches of home games are perceived as units. Since home games are statistically more likely to be won, however, large parts of both groupings may be induced by the season schedule and not indicative of interesting variation in performance.

To eliminate some of the redundancy, and to focus more on performance variation, we could omit expected results  and show only home losses and away wins so that grouping occurs for the performance-induced and relatively surprising results. In this way, the two data dimensions, home-away and win-loss, are no longer represented separately in presence-absence of a horizontal line and above-below center placement of a vertical line.

Observe that horizontal merging of home-game lines is visually dominant. Since home-away schedules in baseball are streaky by design, this may not be the most interesting piece of information, however. In a variant gestaltline, we only place dots on the center line to indicate a win. Using above-below ticks as before, every single outcome can still be uniquely decoded, but groups and gaps on the center line now indicate streaks of wins or losses. We therefore see more easily that Tampa Bay's 2004 season  started out ordinarily with a few too many home losses. They had a long winning streak midseason, including a series of away wins, and after roughly two-thirds of the season there are two particularly poor stretches of home losses.

In comparison to the original sparkline, we have transformed the data to be relative to a baseline (wins at home and losses away) to reduce visual complexity, and deter-

mined glyph parameters from combined data dimensions to place more emphasis on the most important aspect (performance).

Periodicity in Geyser Eruptions

An example of complex repetitive patterns are the eruption sequences of Old Faithful. The geyser, which faithfully erupts about twenty times a day, is a major tourist attraction in Yellowstone National Park, Wyoming, USA. Understanding and predicting the geyser's behavior has been subject to various scientific studies. Among the most recognized ones is the investigation of Azzalini and Bowman [1990] which has been based on 299 successive observation pairs of waiting time between the starts of eruptions (43 to 108 minutes) versus duration of the following eruption (50 to 327 seconds). Both waiting time and duration are bimodal distributed, and the scatterplot in Figure 4.3 reveals three distinct eruption patterns that can be recovered from 3-means clustering or simple thresholding just the same.

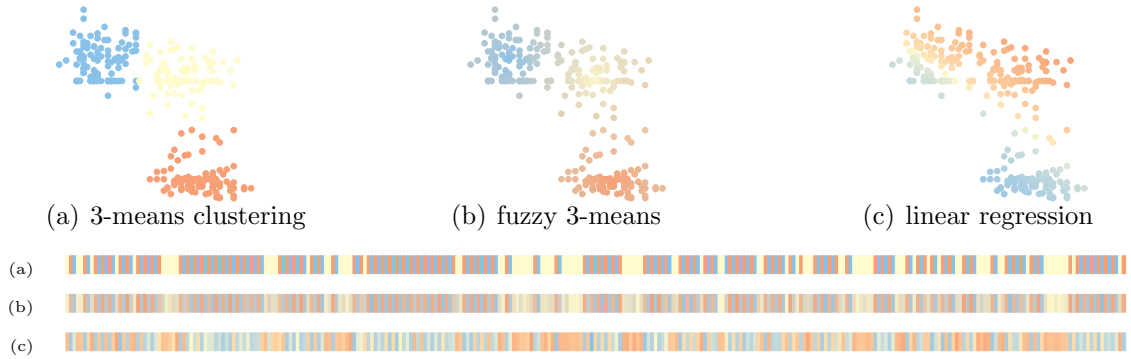


Figure 4.3.: Scatterplots of geyser eruption patterns showing waiting times (x-axis) versus subsequent eruption duration (y-axis) with colors indicating membership in cluster models (a),(b) and estimation error in regression model (c). The gestaltlines below show the actual sequence of eruptions using the same colors.

A gestaltline ... with adjacent stripes colored according to a 3-means clustering model reveals a known periodicity: stretches of an alternating sequence of short waits for long eruptions and long waits for short eruptions are interrupted by shorter stretches of long waits for long eruptions.

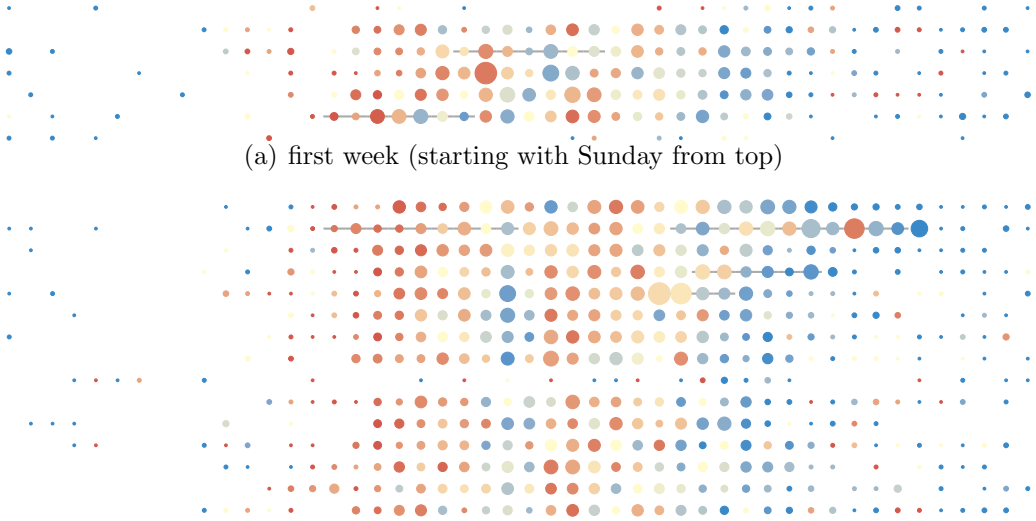
Replacing the partition into groups by fuzzy memberships values, however, yields a gestaltline in which the law of good continuity lets us perceive a continuous alternating pattern ... that is only subdued during what seemed to be interrupts. To the best of our knowledge, this observation is even a new finding not yet reported in the literature.

In comparison to the clustering, the regression model ... appears to yield much less systematic outcomes.

Exceptions in People Flow

Ihler et al. [2006] recorded the number of people going in and out of a building on the University of California at Irvine (UCI) campus over a 15-week period in 2005. Their interest was in modeling relatively stable multilevel (daily, weekly, seasonal) behavioral patterns to detect unusual events.

In their original publication, these data are depicted in standard scatterplots of occupancy or line charts of entry numbers [Ihler et al., 2006, Figures 1 and 3]. Provided an explanatory variable and a period can be fixed in advance, outliers and periodicity are fairly easy to recognize in such diagrams.



(a) first week (starting with Sunday from top)

(b) all Mondays

Figure 4.4.: Net flow of people entering and leaving a building [Ihler et al., 2006, Frank and Asuncion, 2010]. Each row represents a day, each column a 30 minute interval. The area of a dot is proportional to the maximum of the number of people entering and leaving the building within the corresponding half hour, whereas the color indicates the ratio of in- and out-flow on a color scale from red (in) via yellow (balanced) to blue (out). Horizontal background lines indicate known exceptional events taking place in the building.

Now consider the gestaltlines in Figure 4.4 which demonstrates how the recorded data could be shown in full on a single page. Arranging each day in a row of its own supports the detection of weekly patterns such as low building occupancy during weekends, and the vertical alignment of daytime supports the detection of daily patterns such as lunch break times. Unusually late arrivals on Wednesday of the first week (Figure 4.4(a)) and a Monday holiday (Figure 4.4(b)) stick out by breaking perceived groups of similar dots.

So far, we have made similar use of glyph parameters determined from combined data dimensions to direct attention to information derived from the data rather than to the raw data itself.

Some apparent outliers in this data can be related to known special events taking place in the building. An additional horizontal line segment in the background connects the dots inside the time interval corresponding to such events. At least this is how we perceive what might actually be short line segments between neighboring dots. This is because of the law of closure (the segments are aligned and of equal appearance) and the law of figure and ground (consistent gray color). By the law of connectedness, the entire occasion is perceived as a whole and discounted for when eyeing for groups in the remaining data.

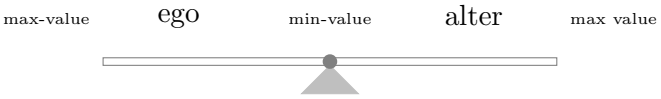
4.2. Gestaltline Design for Longitudinal Network Data

We now turn to the question on how to design gestaltlines for longitudinal network data. More precisely, we assume a sequence of weighted directed graphs that encode a single valued asymmetric relation on a fixed set of actors, rather than, e.g., dyadic events. As we have noted in Section 2.2, network panel data is the standard scenario for the form of analysis that is based on stochastic actor-oriented models.

Initially, we will focus on a basic design for the exploration of individual (asymmetric) dyadic relations (Section 4.2.1). On a second level, then, the alignment of multiple dyad time series gestalts in a matrix arrangement is expected to support exploration of groups of relations (Section 4.2.2).

4.2.1. Seesaw Design: Exploring Individual Dyadic Relations

A simple design, exploiting the impact of gestalt principles and mapping relational/dyadic (ego-alter) data quite naturally to graphical concepts, emanates from the metaphor of a seesaw or scale, respectively. More specifically, take a transparent tube which is subdivided and pivoted in the middle:



Being empty in the beginning it is successively (un)filled from the middle, according to the extent of ratings on a relation

which, of course, may result in imbalance (and and can be interpreted accordingly):

With this rather minimalistic glyph design, including the additional dimension of time to form a single (letter-like) dataword is now straightforward. Some may find the stacking below reminiscent of a modernistic depiction of a falling tube tumbling left and right according the balance of its filling. This is intentional because it is in line with our basic metaphor. Since human perception is geared toward comparison of progress with respect to some grounded baseline, however, we use bottom-up visualizations of development instead; compare Figure 4.5.

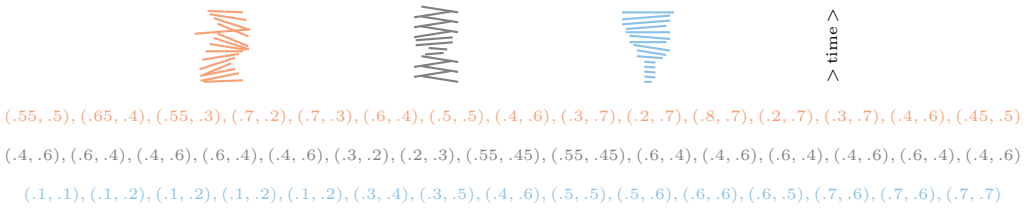


Figure 4.5.: Seesaw gestaltline design (read bottom to top) for dyadic time-series (printed in the same color below; read left to right) such as ego-alter ratings on a relationship. Patterns, trends, and outliers are clearly visible and interpretable — "Their relationship was an emotional seesaw."

The graphical mapping displays both quantities of interest, the absolute values (extent) and their differences (balance/reciprocity), simultaneously:

- the higher the weight on an edge, the more ink
  (irrespective of the degree of balance)
- the more imbalance in a dyad, the more prominent the slope
  (irrespective of the size of weights)

Our design choices are based on intuition and gestalt principles. While a single seesaw intuitively represents the concept of imbalance, their alignment is intended to invoke the gestalt principle of *common fate*, which states that graphical elements of similar orientation are perceived as a whole even when they differ in other attributes such as size.

As we have argued in Section 4.1, such a conscious exploitation of gestalt principles appears to add a new twist to the use of glyphs. This may contrast the proposed graphical mapping from other designs working also in small-scale such as juxtaposed bar charts mapping time on the vertical axis (cf. Figure 4.6), (rotated) horizon graphs [Heer et al., 2009] or data vases [Thakur and Rhyne, 2009].

Because of their small, letter-like space requirements, individual dyad gestalts can be used directly inside of (publishable) text. In this way, it is not necessary to go back and forth between an explanation and a (network) diagram or animation. For illustration compare the two paragraphs in Figure 4.7 with regard to the entropy of information: The use of seesaw gestaltlines provides a means to write with data and "see" the entire dyadic evolution, i.e. it is facilitating comprehension, improving readability and conveying more information. Therefore, the proposed *static* gestalt-based

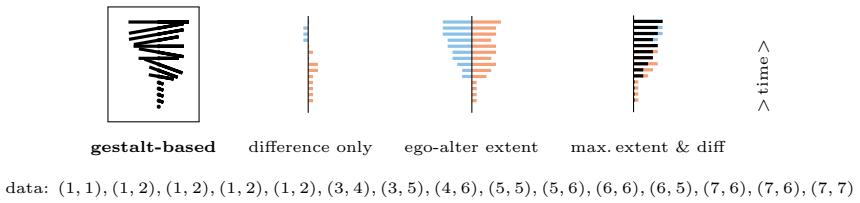


Figure 4.6.: Gestalt-based representation of dyad evolution compared to conventional times series charts displaying the difference of values, contrasting both values, or juxtaposing maximum extent and difference.

representation provides a convenient and intuitive way to explore and communicate extent and balance in *dynamic* asymmetric relations.

For example, 7  12 are friends almost from the beginning. While 2  befriends 17 only late, 16  17 seems to stand no chance at that. The relationships of 7  14, 3 15, or 10 15 are rather more complicated. The latter is no surprise as 10 is unpopular  with many.

"Nodes 10 and 15, for example, quickly emerge as nodes on the edge of the social structure. While they nominate each other symmetrically early in the observation period, they lose interest in each other by the end. Neither node receives top-five nominations from any other node in the network. Their nominations to others seem to dance around the graph, never resting for long on a single person." [Moody et al., 2005, p. 1228]

Figure 4.7.: Exploration and explanation of individual dyadic relationships with gestaltlines: more information with fewer (data)words. (Newcomb fraternity data)

Moreover, the minimalist nature of the basic glyph design allows for additional enhancements, inspired by gestalt laws. First of all, the current design leaves color as a degree of freedom to represent additional attributes; cf. Harrower and Brewer [2003] for various coloring schemes. For instance, single data values or special data ranges can be highlighted or mitigated, exploiting the law of figure and ground. Another particular application in social networks would be the fading out of unconfirmed relations because these may be a critical factor in relational data collection [Kossinets, 2006]. Alternatively, color could be used to combine and distinguish different ranges of extent and/or balance as suggested by the law of similarity. We will address all these issues in the case studies provided in Section 4.3.1 and Section 4.3.2, respectively.

We already alluded to claim that human perception subconsciously relates progress with respect to some baseline. One may capitalize on this by explicitly indicating a reference line that incorporates additional (statistical) information at the bottom of each gestalt. We even suggest to use two reference lines where appropriate, e.g., one indicating the expected time-series values in the beginning and one indicating the

expected values at the end of the evolution.

Finally, although this is rarely the case in network survey data, we do note that if there are too many points in time, a horizontal representation of evolution might be more appropriate for word-like representations. That is, line segments are depicted next to each other, metaphorically describing a dyadic relationships as being balanced |, or toppling towards ego \ or alter /, respectively; refer to Appendix A for a similar gestaltline design in a different setting.

4.2.2. Gestaltmatrix Design: Exploring Groups of Relations

We have just demonstrated the practical effect of gestaltline visualizations for individual dyadic relationships. Moreover, their character-like representation enables a straightforward integration of dyad gestalts into a matrix-like diagram to enable visual exploration of patterns, trends, and outliers in groups of relations. As a motivating example, in Figure 4.8 we stress once more the potential to complement and critically inform existing analyses, based on animation or interaction.

"For example, one can see that nodes 1, 6, 8, and 13 remain strongly connected to each other throughout the observation period, occupying a small cluster at the right of the graph. Nodes 7, 12, and 4 form a cluster early in the group's history, but node 4 then breaks with this group at about week 8, instead nominating nodes 17 and 2."

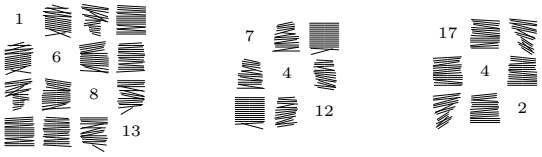


Figure 4.8.: Illustration with combination of gestaltlines casting doubt on statements from the literature [Moody et al., 2005, p. 1228]. (Newcomb fraternity data)

We refer to such a tabular arrangements of seesaw gestaltlines as a *gestaltmatrix*. Each matrix entry y_{ij} depicts a single dyad gestalt, i.e. the evolution of the entire dyadic time-series of ego i and alter j. Note that matrix entries y_{ij} and y_{ji} are redundant, since depicted gestalts are just mirrored at the vertical axes. The full matrix is displayed nevertheless to capitalize on the alignment across rows and columns. It facilitates, therefore, the detection of overall patterns, trends, and outliers from an ego point of view (the row) or from the view of all others (the column), respectively.

Examples of gestaltmatrix representations that convey the entire network evolution are provided in the case studies below. As noted in Section 2.1.2, the permutation of rows and columns is crucial to highlight higher level organization in matrix-based network representations. In the absence of external information on suitable actor permutations, therefore, we describe a dyad gestalt by some similarity measure s that summarizes the time-series information $y_{ij} = s\big((y_{ij}^{(1)}, y_{ji}^{(1)}), \ldots, (y_{ij}^{(T)}, y_{ji}^{(T)})\big)$. Correspondingly defined aggregated actor similarities, then, can be used as the input to a

blockmodeling approach (cf. Section 2.1.2).

The quality of an ordering obtained from such a similarity depends on its suitability for the analytic perspective. As a default that can be used in the absence of more specific requirements, we propose to sum the geometric means of dyadic ratings, $y_{ij} = \sum_{t=1}^{T} \sqrt{y_{ij}^{(t)} \cdot y_{ji}^{(t)}}$. This represents a reasonable general-purpose solution because of its tendency to group actors with consistently high and balanced weights. Part of the rationale is the correspondence of the geometric mean with the amount of ink and slope used in the seesaw design. Note, however, that blockmodeling can be misled dramatically by large heterogeneity in terms of node degrees [Karrer and Newman, 2011].

4.3. Case Studies

With the increasing use of sparklines in popular media such as newspapers,[1] gestaltlines in general might appeal to a broader audience. The special gestaltmatrix design for longitudinal network data, on the other hand, is much more targeted at domain experts and scientific communication. Therefore, in this section, we showcase the intended usage of gestaltmatrix representations and report empirical findings on the added value toward a better understanding of evolving social networks. Examples range from new insights on old data from the sixties (Section 4.3.1) to the usage of gestaltmatrices within ongoing cooperations in the present (Section 4.3.2).

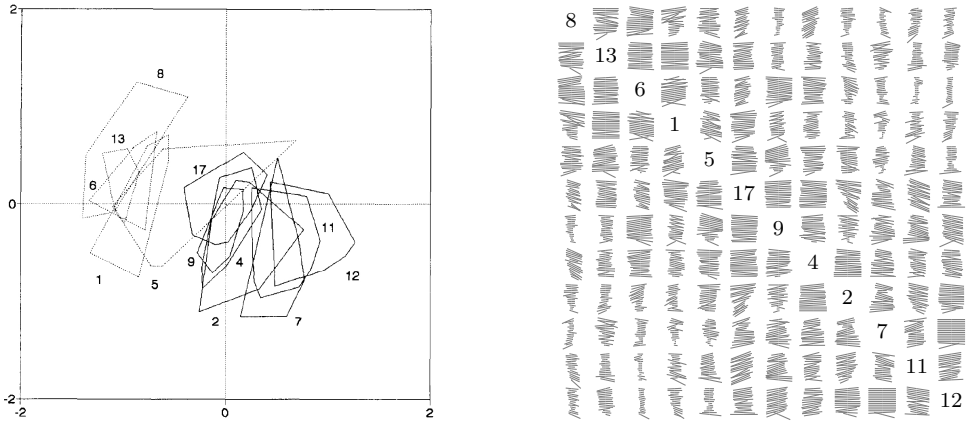
4.3.1. New Insights on Old Data

Initially, we illustrate and further elaborate the proposed principles on one of the most well-known longitudinal data sets in social network analysis, the Newcomb fraternity data (cf. Section 2.2.1). Throughout the observation span, most mutual rankings in the Newcomb fraternity data are highly inconsistent, which indicates a low level of reciprocity or a high level of asymmetry, respectively. Moreover, being the subject of many previous analyses, the data is ideally suited to demonstrate our method's capability to provide previously unexplored insights.

In particular, the evolution of group structures within the Newcomb fraternity data has been studied extensively. Evaluating blockmodeling algorithms according to the density of inter- and intra-block top ratings, e.g., White et al. [1976, page 764] state that "by at least the fifth week not only the final blocks but also the final blockmodel have emerged with remarkable clarity". Their findings basically confirm the actor groups originally proposed by Nordlie [1958] based on rank correlations. Nakao and Romney [1993] provide further evidence for structural convergence in the Newcomb fraternity data by relating the number of concordant (c) and discordant (d) rankings with Goodman-Kruskal's Gamma coefficient $(c - d)/(c + d) \in [-1, 1]$.

[1] Refer to `http://www.edwardtufte.com` for examples and discussions on the use of sparklines, e.g., in Microsoft Excel.

Findings on group structures in the Newcomb fraternity data have been presented either textually (13 9 17 1 8 6 4) (7 11 12 2) (14 3 10 16 5 15) [White et al., 1976, page 764] or by depicting a permuted sociomatrix for each wave; compare Figure 2.1(b). An exception can be found in Nakao and Romney [1993], where the authors use multidimensional scaling to place actors in the plane according to the similarities of their rankings at a given point in time. Then, Procrustes analysis is used to align each actor's positions over time, as indicated by the convex hulls presented in Figure 4.9(a). Any of these previously detected actor groupings can be compactly visualized and evaluated in a single gestaltmatrix without data aggregation; compare Figure 4.9(b).



(a) [Nakao and Romney, 1993, p. 122] (b) Gestaltmatrix

Figure 4.9.: Previous and new description of groups in the Newcomb fraternity data.

The gestaltmatrix example in Figure 4.9(b) is reassuring for a subdivision into two primary subgroups as suggested by the distinction of dotted and solid polygons in 4.9(a). Yet, using the same amount of space, much more detailed information and additional insights on the strength of cohesion over time can be gathered from the gestaltmatrix visualization. For instance, actor 2's friendship efforts towards (popular) actor 17 are reciprocated only late, while 7's efforts towards 17 become less successful in the end.

Previous studies on the Newcomb fraternity data have also noted that the persistent asymmetry of friendship rankings comes along with a large heterogeneity in the popularity of actors. As suggested in Section 4.2.1, we will now augment the plain color design to highlight this additional information on the network evolution. First, inspired by the gestalt principle of *figure and ground*, we combine two different hues: using black for highlighting top 3 values of the extent and gray for fading the others, we intend to assess the result of a commonly applied thresholding before the background of the full dataset; cf. Section 2.2.1. Second, motivated by the gestalt principle of *similarity*, we distinguish different types of actors by coloring them according to their standardized popularity over time. Popular actors are colored blue, whereas unpopular actors are colored red.

Figure 4.10.: Gestaltmatrix of Newcomb fraternity data, showing the complete evolution of 4080 rankings collected in 15 waves. Labels on the diagonal are numbered according to Nordlie [1958] and colored according to standardized overall popularity of corresponding actor. The arrangement of actors is according to the sorting proposed in Nakao and Romney [1993].

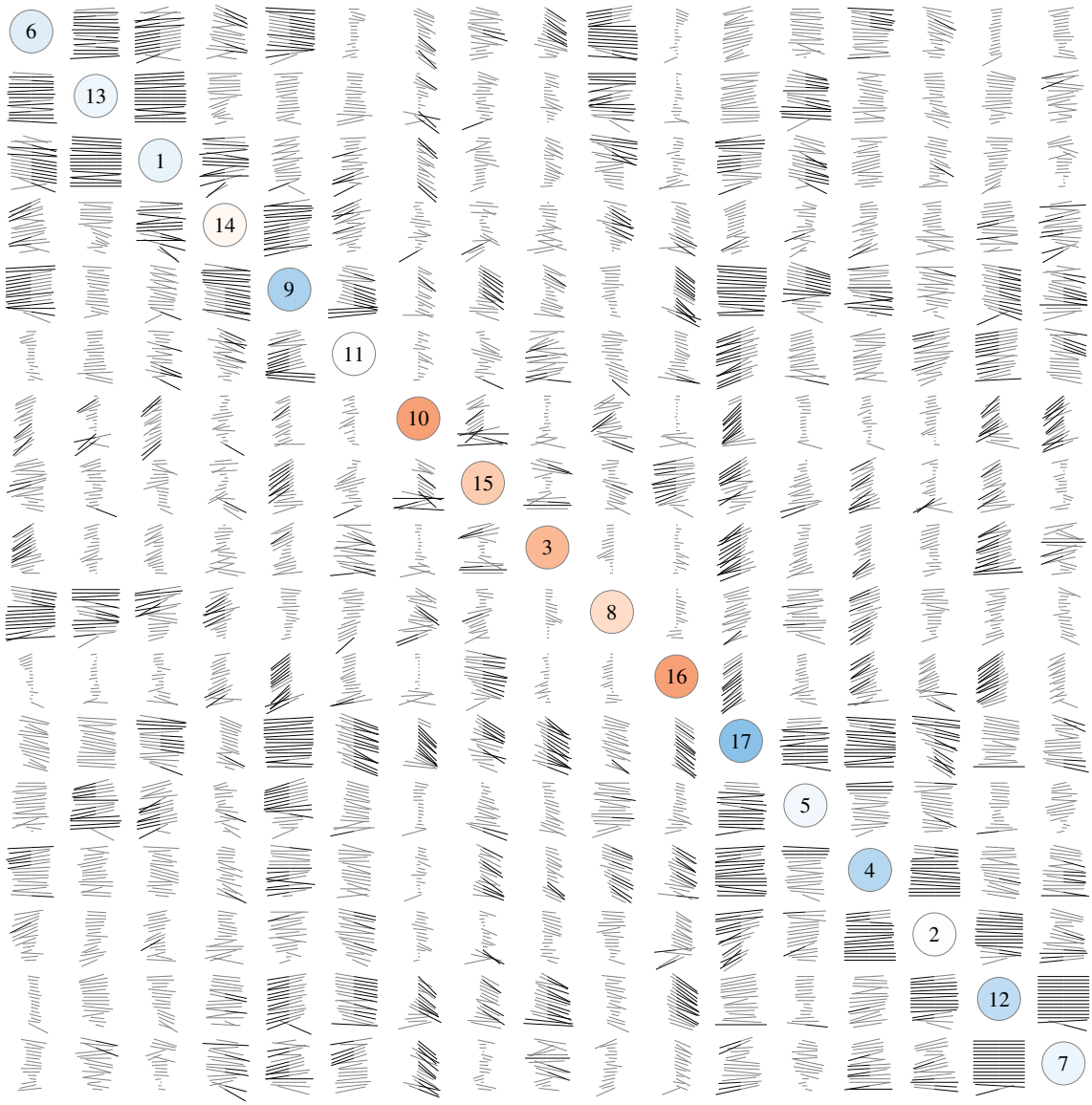


Figure 4.11.: Gestaltmatrix of Newcomb fraternity data, showing the complete evolution of 4080 rankings collected in 15 waves. Labels on the diagonal are numbered according to Nordlie [1958] and colored according to standardized overall popularity of corresponding actor. The arrangement of actors is according to a spectral approach described in the main text.

More specifically, we evaluated the deviation π from a null model, assuming that at each time step t, each actor i obtains each possible ranking at a time, i.e. $\mu = \sum_{j \neq i} y_{ji}^t / k = k/2$, with $k = \#\{j \neq i\}$; assuming rankings from k to 1. This gives

$$\pi(i) = \frac{1}{T} \cdot \sum_{t=1}^{T} \sum_{j \neq i} \frac{y_{ji}^t / k - \mu}{\sigma}$$

where σ refers to the standard deviation of a single ranking y_{ji}^t from μ, and T is the number of network observations.

A gestaltmatrix of the Newcomb fraternity data with interpolated colors according to this deviation is shown in Figure 4.10. In this matrix we reproduce an ordering of all actors as proposed in Nakao and Romney [1993]. Yet, again, with the gestaltmatrix view on the complete evolution of relations, additional insights can be obtained from the more refined perspective. For instance, consider the column of the "scapegoat in this group (man 10), who received one of the bottom three choices of each of the other persons" [White et al., 1976, p. 759]. It is clearly visible which others did overrate the 'scapegoat' in the beginning. Interestingly, these include the overall most popular actors 17  , 4 , 9  and 12  . Also, a strong desire of the most unpopular actors to be friends especially with the popular actors is revealed. To the best of our knowledge, these findings on the Newcomb fraternity data have never been published before, despite long-standing investigations spanning various disciplines.

As a final variant gestaltmatrix of the Newcomb fraternity data with identical coloring, we propose a new actor ordering based on the number of inversions in the rankings of consecutive weeks. The corresponding line chart  assumes local minima in week 8 and week 12. Thus, the ratings in week 7 and week 11 are rather stable and, therefore, reasonable proxies for static network representations. We choose week 11 and sort actors according to their values in the Fiedler vector of the Laplacian matrix obtained from $(y_{ij})_{i,j}$ where $y_{ij} = \sqrt{y_{ij}^{11} \cdot y_{ji}^{11}}$ if both ratings are top 3-ratings and $y_{ij} = 0$ otherwise. This choice of similarity measure, again, is motivated by the common use of thresholding in published reanalyses of the Newcomb fraternity data; compare Figures 2.3 and 2.4. The resulting gestaltmatrix is presented in Figure 4.11. A block structure due to higher internal rankings is clearly visible.

4.3.2. Gestaltmatrix Visualizations in Practice

The last sections have demonstrated the intended usage of gestaltmatrices for scientific communication. In this section we provide two additional examples of interdisciplinary cooperations in which gestaltmatrix representations have been welcomed with open arms by the corresponding domain experts.[2] Since most of that data is proprietary and has not been published yet, however, the following instantiations of gestaltmatrix representations are meant as an exemplary illustration of the relative merit of our

[2]The author gratefully acknowledges the permissions of Francesca Pallotti and Britta Renner to showcase the two gestaltmatrix visualizations of their data.

ideas rather than a genuine analysis of the underlying data. Still, the two case studies shed light on additional aspects of gestaltmatrix visualizations that are relevant for the intended application to longitudinal network survey data, namely, the scalability to anticipated data dimensions and a capability to cope with missing information [Marsden, 1990, Kossinets, 2006].

Revealing Core-Periphery Structures in Hospital Patient Transfer

As a prime example for relational collaboration among spatial multipoint competitors, Lomi and Pallotti [2012] investigate inter-hospital patient transfer in Lazio, one of the 20 administrative regions of Italy. The relation of patient transfer among hospitals is highly asymmetric, in the sense that the number of patients sent from one hospital to another typically differs much from the number of patients that is received in return.

While the analysis in Lomi and Pallotti [2012] is based on a single network observation of patient transfer relations among 91 hospitals in the year 2003, the same authors have provided us with not yet published longitudinal data that was collected in the same region. In detail, the data consists of five cooperation matrices recording the annual number of patients that have been transfered among a set of 110 hospitals in the years from 2004 to 2008. From the domain experts point of view, given the time series of asymmetric weighted graphs, the primary interest is to visually emphasize and explore the strength and direction of patient flows.

The data clearly constitutes an interesting case for analyzing asymmetric relations in longitudinal networks by means of gestaltmatrix visualizations and differs significantly from that of the Newcomb fraternity data. In particular, since the majority of empirical longitudinal studies typically deals with networks of 10 to 100 actors (cf. Section 2.2), the number of dyadic relations is at the opposing end of the anticipated data dimensions.

We use the following instantiation. According to the general gestaltmatrix design, each gestaltmatrix cell depicts the extent of patient transfer over time (from bottom to top). In a sense reversed to the previous examples, however, the *in*-transfer is understood as ego's "rating" and the *out*-transfer is understood as alter's "rating". This decision is motivated by resulting imbalances that can be metaphorically interpreted as patients "sliding" towards a hospital (more in-transfer) or rather away (more out-transfer). Moreover, we use a very dominant graphical attribute, color, to distinguish the different quality of more in transfer, more out transfer and exactly balanced patient transfer.

Not only the dimension of the underlying edge-weighted directed graphs significantly differs from the previous examples on the Newcomb fraternity data, but also the valued asymmetric relation itself is much more complex: Actual (non-zero) patient transfer on the dyadic level ranges from 1 to 810 patients a year, and the distribution of values is strongly positively skewed (mean transfer per year is 8.04; standard deviation 24.8). As a consequence, we applied sublinear scalings to determine the length and slope of

Figure 4.12.: Gestaltmatrix representation revealing local core-periphery structures in the annual patient transfer relations within a regional community of 110 hospitals. Matrix partitioning and hospital arrangement according to externally defined units, coloring according to the aggregated balance of in- and out-transfer (zoom in to see more detail). *Data courtesy of Francesca Pallotti.*

line segments.[3]

In line with the previous design decisions, we use the identifying labels (on the diagonal of the gestaltmatrix) to visualize additional summary information on each hospital: We emphasize the size of each hospital's label according to the overall extent of (in- and out-) patient transfer and select a coloring according to the corresponding balance of in- and out-transfer.

Finally, the data contains exogenous geographic information at the hospital level, which allows for attribute-based enhancements of the basic gestaltmatrix design. In particular, each hospital is uniquely assigned to one of 12 *local health units*. We use this information to arrange the hospitals accordingly and introduce auxiliary lines that visually subdivide the matrix into the blocks that correspond to the basic partitioning of the regional health system in Lazio — a second dyadic covariate, encoding the geographical distance between hospitals, is used to sort the blocks according to the mean distance of inter-block hospital distances. Moreover, the intra block arrangement of hospitals is determined according to the extent of intra local health unit patient transfer.

As a result of these design choices, for the first time, the domain experts were able to "see" and interpret their data it is entire form; cf. Figure 4.12. For example, the domain experts were left with the ability to visually detect exceptional dyads, such as the remarkable periodic patient exchange ≷ between 066 and 058. Moreover, the gestaltmatrix representation was not only emphasizing sources and sinks in the network of patient transfer (as expected), but also revealed striking local core-periphery structures, with cores emerging in the upper left corner and periphery nodes emerging in the lower left corner of each block on the diagonal. This enabled the domain experts to explore the strength and direction of patient flows with regard to an additional quality that would have been difficult to arrive at analytically. Clearly, these visual encounters entail further substantive work beyond the scope of this illustration.

Compensating for Missing Data in Evolving Friendship Networks

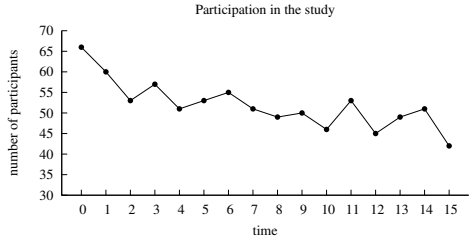
As a second application of gestaltmatrix visualizations in practice, we touch on data from the "SozNet" study on social networks and health related questions – such as perceived and actual social discrimination [Hartung and Renner, 2013] – conducted at the University of Konstanz.

The small part of the SozNet data we report here was collected during a study on an evolving friendship network among 78 university freshmen of the psychology department. Once a week, students filled out a questionnaire, by which personal attributes, contact information and, possibly, friendship relations were assessed. In total, there were 16 observation time points.

[3] Given the number of patients y_{ij}^t being sent from hospital i to hospital j in year t, the length of each line segment is drawn proportional to $\sqrt{\log\left(1+y_{ij}^t\right)}+\sqrt{\log\left(1+y_{ji}^t\right)}$, and the slope of lines is selected proportional to $\operatorname{sgn}(y_{ij}^t - y_{ji}^t)\cdot\sqrt{\log\left(1+|y_{ij}^t - y_{ji}^t|\right)}$.

The data at hand contains a considerable amount of missing information. Firstly, participation in the study ranged from 66 students in the first week to 42 students in the last week. While 24 students participated in all measurements, the majority of students participated only irregularly, and 12 students even never participated at all. Figure 4.13(a) shows the number of participants in all observations. Secondly, the social network of the freshmen was collected by first asking each student whether there had been contact with other students. Only if contact was confirmed for a given dyad and week, detailed friendship information between the questioned person and a given fellow student were assessed.

For the purpose of this illustration we only reproduce friendship information on how good the persons *know* each other together with the gender covariate. The resulting `knowing` network information is still expressive, since, with one exception, all relational attributes in the original data are highly correlated to each other; cf. Table 4.13(b).[4]



(a) Number of participants per week

days of contact	place of contact	intimacy	knowing person	liking person	give support	get support
1.00	0.52	0.60	0.64	0.48	0.44	0.45
	1.00	0.34	0.36	0.21	0.17	0.18
		1.00	0.64	0.54	0.48	0.50
			1.00	0.58	0.50	0.53
				1.00	0.80	0.76
					1.00	0.88
						1.00

(b) Average covariance of relational attributes.

Figure 4.13.: SozNet study on an evolving social network of 78 university freshmen.

Given the large amount of missing values, the SozNet data exhibits another quality than the complete data sets we have seen before — the more so as relationships in the `knowing` network are not as asymmetric as the relations in the examples we have illustrated before. In this respect, Figure 4.14 is designed to illustrate that, to a certain degree, gestaltmatrix visualization are able to complement the treatment of non-response in longitudinal network studies [Huisman and Steglich, 2008] by way of visually compensating for missing values rather than imputing such information analytically [Huisman, 2009].[5]

In particular, since the survey design conditioned the measurement of relational attributes on weekly encounters, we are left with occasional white space in the stacked seesaw design of the entire dyadic evolution which is rather random. Still, by the Law of Prägnanz, a viewer is perceiving the stacked seesaw gestalt as a whole. Therefore, one can visually compensate for a certain amount of breaks, by organizing the disparate visual stimuli of individual weekly ratings and occasional white space into the simplest stable and coherent form. The mind will make an educated guess for

[4]Since the collected data is proprietary, we do not provide details on any other attributes.

[5]In an analytical study similar to the one reported in Chapter 5, for instance, we decided to replace missing information in evolving friendship networks with the most recent information available in previous observations [Brandes et al., 2009a].

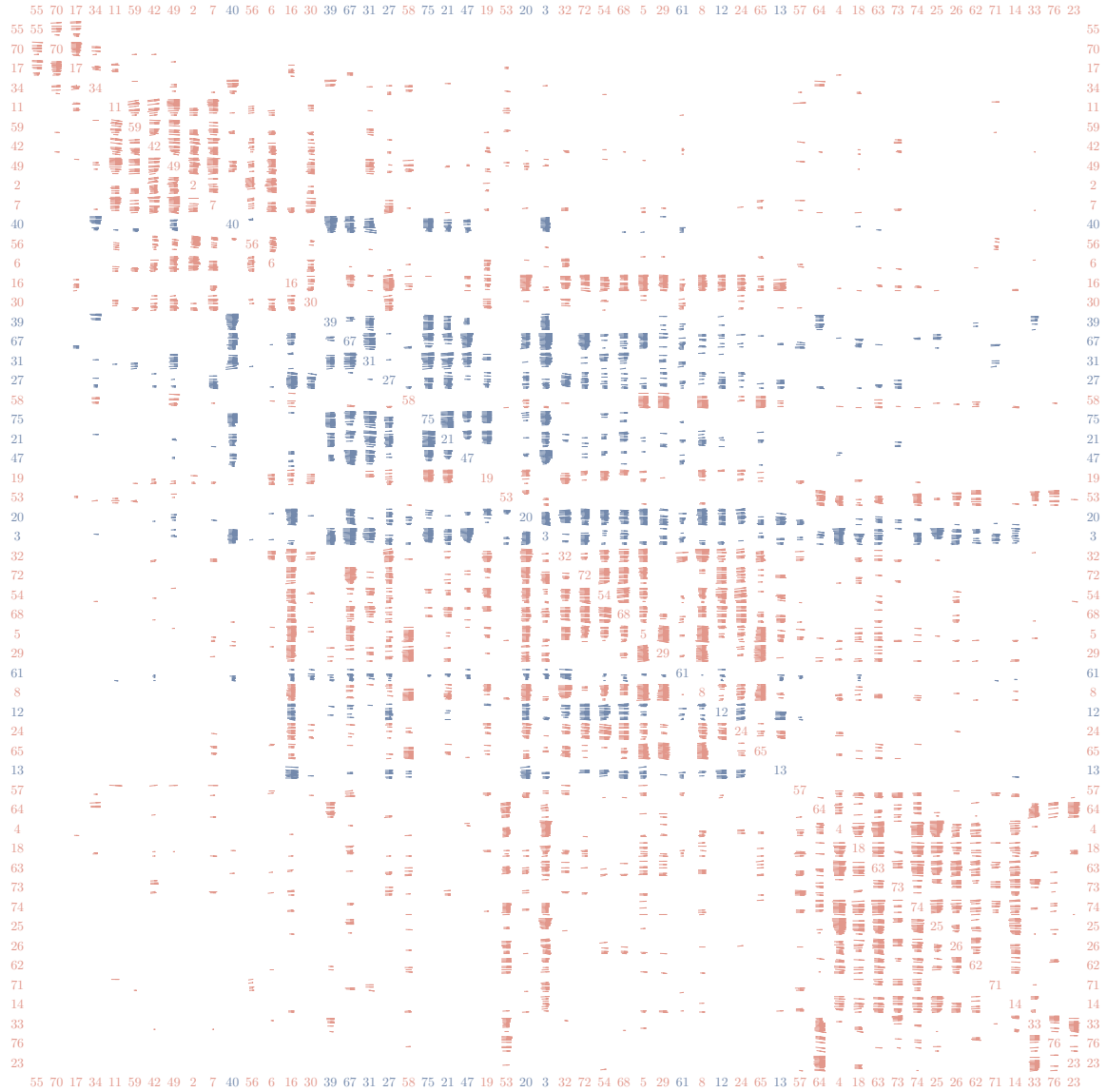


Figure 4.14.: Gestaltmatrix representation compensating for missing data in an evolving friendship network among 78 university freshmen, with color indicating gender. *Data courtesy of Britta Renner.*

missing values, which can be best explained by the law of closure and the law of good continuity (cf. Figure 4.2). Moreover, using the default spectral approach to define a suitable permutation, the gestaltmatrix representation provides a longer-term picture of seemingly unstable relationships as being part of stable groups of boys (cyan) and girls (magenta). Also note the cross-like patterns resulting from popular actors. Both these visual groupings are not altered by missing values significantly.

We are happy to report that the domain experts felt similarly confident with this alternative representation of their data.[6]

4.3.3. Expert Feedback

We received very positive feedback both from our collaborators and unbiased experts in information visualization: "the technique seems like a very intuitive way of viewing relationships".[7]

We will ultimately be interested in a quantitative comparison assessing the relative efficiency of our design in communicating extent and balance in asymmetric relations. In the absence of sufficiently similar previous approaches, however, the number of dimensions and confounding factors appears to be too large for a formal, well-focused user study at this place [Kosara et al., 2003]. Instead we rely on external expert reviews as an initial sanity check [Tory and Moller, 2005].

Because of its intuitiveness and some other obvious benefits (static representation, generality, printability, publishability) we did assume that social scientists are willing to spend the five minutes it takes to understand the design and learn how to read it. The method was therefore presented at INSNA Sunbelt XXXI Social Networks Conference,[8] the main venue for networks in the social sciences (no published proceedings). Interest was surprisingly high and feedback overwhelmingly positive; even to the point that the idea was taken up and applied in another presentation at the same conference within two days.

Fourteen senior domain experts and a number of other delegates provided personal feedback in private and informal face-to-face discussions, with all but three approaching us before we could ask them to. They mentioned in particular the simplicity, intuitiveness, and aesthetic appeal of the design, and the ease by which tendencies and outliers can be detected both on the actor and dyad level. They also mentioned the compact and simultaneous representation of the entire data set, allowing back and forth comparison of matrix rows and cells. Generally, they were aware of the fact that distances between vertices are difficult to determine in matrix representation, but needed help in understanding the consequences of row and column reordering and the reliability of conclusions drawn from a given ordering. The latter was somewhat

[6] "Die grafische Darstellung der SozNet Daten ist genial." (German; the graphical representation of the SozNet data is brilliant; extract from email correspondence with domain expert not mentioned by name)

[7] http://infosthetics.com/archives/2011/11/most_interesting_papers_at_infovis_visweek_2011.html

[8] 8-13 February 2011, St. Pete Beach, FL, USA; approx. 550 delegates.

surprising because this aspect of our design is shared with any matrix representation of networks (cf. Section 2.1.2).

As we had hoped for, feedback was particularly vocal from the modeling community. The gestalt-based design was found to serve the purpose of visually detecting evolving patterns and, in particular, exceptional actors, better than anything our respondents had known before.

In the absence of a quantitative assessment, we feel that at the very least this anecdotal evidence proves domain relevance. Moreover, again, the interested reader is also referred to Appendix A, in which we provide empirical validation of gestaltline visualizations in a non-network setting.

4.4. Discussion

We proposed a novel methodology for static visualization of longitudinal asymmetric network data, which we refer to as *gestaltmatrix*. It is based on word-sized representations of dyadic evolution, which in turn represent an application of gestaltlines, i.e. multivariate sparklines capitalizing on gestalt principles.

To the best of our knowledge, the most similar approach to representing change in social networks has been proposed in Stein et al. [2010]. There, the authors indicate time through various subdivisions of matrix cells, mapping each data value to a single colored pixel. Such an atomistic treatment is highly efficient in terms of visualizing "the largest amount of data which is possible on current displays" [Keim, 2000]. In terms of gestalt laws, however, a holistic mapping seems to be more appropriate for indicating extent and balance simultaneously; cf. Figure 4.6. Other related approaches involve animation and interaction, such as toggling the colors of matrix cells according to a queried time range [Ghoniem et al., 2004].

While, currently, scepticism with regard to glyph-based approaches to representing multidimensional data appears to dominate, our work is in line with other recent work such as Nielsen et al. [2009], in which it is suggested that additional perceptually-effective forms of compact multidimensional representation may await discovery and characterization.

Since our approach provides novel means to explore and communicate the extent and balance of values in dyads, it can also be used to complement existing techniques that require animation or interaction. Next to some obvious benefits such as effective communication of findings and suitability for publication on paper we also demonstrated that gestalt-based network analysis bears the potential to yield additional insight even into data that was previously studied extensively. Additionally, we revealed misinterpretations that we suspect resulted from more aggregate data views.

The provided case studies have demonstrated that there are many ways in which additional information about endogenous and exogenous variables can be integrated into a gestaltmatrix network visualization. The main goal was to argue that a detailed design can lead to interpretable forms on the level of the dyad (matrix cell), the actor (row/column), and the network (matrix). These forms are likely to ease the discovery

of trends, change events, and outliers.

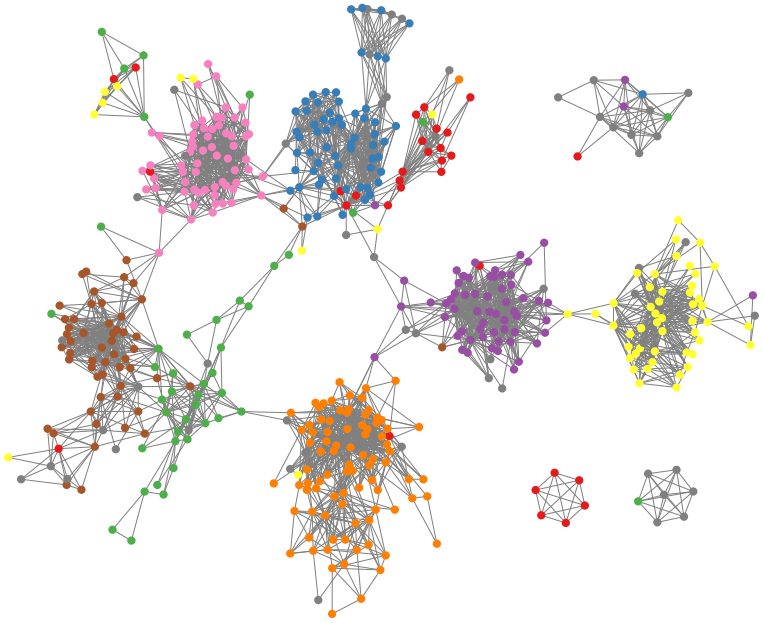
There are two major limitations for the scope of our method. The first one is shared with all matrix representations of networks, namely that paths are difficult to discover and follow. The other one is a consequence of our attempt to bring out a joint appearance of the data in a dyad; as a consequence, individual time slices are difficult to extract. Possible remedies for these two problems may lie in a combination of the seesaw metaphor with node-link diagrams and by using gestaltlines highlighting the current point in time as edge labels. These, however, require more careful research. The case studies in Section 4.3.2 can be considered an upper limit for the large majority of empirical studies also from the modeling point of view. For larger networks, however, larger print and, possibly, a hierarchical design may become necessary.

Qualitative evidence on the acceptance and intuitive understanding of the proposed principles was provided based on informal feedback from domain experts. Further research shall include a quantitative assessment of the impact that gestaltmatrices have on the understanding of asymmetric relations in longitudinal social networks. A first step in that direction (regarding the justification of gestaltlines in general) is reported in Appendix A.

Chapter 5.

Conditionally Independent Tie Changes

5



The findings of this chapter have been previously published in the *Journal of Mathematical Psychology* [Lerner et al., 2013].

As we have pointed out in Chapter 3, mutual dependence among dyads is one of the most universal characteristic of social network data. Yet, empirical network data often contain some – but not perfect – time information about the ordering and timing of dyadic evolution (cf. Chapter 2). This availability of temporal information adds another dimension to the notion of dependency: it is reasonable to assume that past network evolution is *not* influenced by future network evolution. Therefore, to a certain degree, mutual dependencies of tie changes might be excluded from longitudinal network models, if a preceding network configuration is taken into account. The goal of this chapter is to shed light on possible answers to this question.

There are several sources of motivation for conducting this study. First of all, we observe that conditional independence of tie changes is often (sometimes implicitly) assumed in published research; examples include Crescenzi [2007], Maoz et al. [2007], and Hanneke et al. [2010] for the case of panel data about relational states — as well as Brandes et al. [2009c] and Stadtfeld [2012] for the case of interval-censored relational event data (cf. Section 2.3.2). It is thus relevant to test the validity of this assumption and to tackle the question whether it affects the reported findings. Moreover, answering the question to which degree tie changes mutually depend on each other for given data is of high practical relevance since, *if* the danger of drawing invalid conclusions can be ruled out, conditional independence models are computationally much simpler and lead to more stable estimation algorithms. Finally, as a rather general and intrinsic argument, we simply want to learn as much as possible about the dependence structure in empirical social network data.

Consequently, we propose a framework to systematically compare simple network models that assume conditional independence of tie changes with more involved models that do not. Furthermore, we treat the question whether substantive findings about hypothetical network effects, such as transitivity, are affected by the (potentially invalid) assumption of conditional independence.

Note that, e.g., Steglich et al. [2010] argue against models that assume independence of tie-change events if a sequence of network evolution steps is only partially known. It is not our goal to challenge this claim but rather to quantify the implications of assuming independence when analyzing given longitudinal network data. While it can be expected that conditional independence models are inappropriate as a *general* model for network evolution, it is possible that the independence assumption becomes less severe when the observation intervals are relatively short compared to the usual rate of change of the analyzed relation.

5.1. Framework of Analysis

As an introductory example, compare different change scenarios that ultimately result in a transitive triangle  configuration: a longitudinal network model that assumes conditional independence of tie-change events will not recognize an evolution starting from an initial single tie as transitivity effect, since, e.g., a third tie closing a

directed two-path  that was not present in the beginning, can only be taken into account by more complicated conditional dependence models.

We propose a framework to systematically compare conditional independence models with more general models that are specifically designed for social network data. For clarity of exposition, the following illustration is focused on panel data about networks of relational states. Still, in Section 5.1.3 we will outline that interval-censored relational event data can be treated with very similar methods.

5.1.1. Conditional Independence and Lagged Network Statistics

Given multiple observations of the network evolution, i.e. network panel data, we assume the corresponding adjacency matrices to be realizations of a continuous-time random process $Y^{(t)} \in \{0,1\}^{n\times n}$, $t \in \mathbb{R}$, observed at time points $t_1 < \cdots < t_T$; cf. Section 2.2.

Tie changes that happen in the interval (t_{h-1}, t_h) typically depend on the preceding networks at $t_1, \ldots, t_{h-1}$. The *conditional independence assumption* states that, for all $h \geq 2$, the conditional distribution of the network $y^{(h)}$, given the networks $y^{(1)}, \ldots, y^{(h-1)}$, is the product of the conditional probabilities of the individual ties. Formally, for all $h \geq 2$ it is

$$P(Y^{(t_h)} = y^{(h)} | y^{(1)}, \ldots, y^{(h-1)}) = \prod_{i \neq j} P(Y_{ij}^{(t_h)} = y_{ij}^{(h)} | y^{(1)}, \ldots, y^{(h-1)}) \ . \tag{5.1}$$

Thus, in a conditional independence model, the probability of a tie between i and j at time t_h might depend on the presence or absence of another tie, say (i,k), at time points $t_{h'}$, $h' < h$, but not on whether the tie on (i,k) was created or dissolved in the interval (t_{h-1}, t_h).

In Chapter 2 we introduced the notion of actor covariates and dyadic covariates. As a basic example, models that only include constant covariates as explanatory variables clearly satisfy the conditional independence assumption. To allow for models that include dependence among dyads but still satisfy the conditional independence assumption, we introduce another class of network statistics resembling dyadic covariates but having a fundamentally different interpretation.

A *lagged (dyadic) network statistic* is defined to be a family of $T-1$ real $n \times n$ matrices $x^{(1)}, \ldots, x^{(T-1)}$ that are deterministic functions of the observed networks $y^{(1)}, \ldots, y^{(T-1)}$, respectively. Similarly to covariates, the lagged network statistics are kept constant in the intervals between observations $[t_{h-1}, t_h)$ and are assumed to influence the evolution of ties within these intervals. For example, a lagged network statistic at time t_{h-1} might encode for each pair of actors (i,j) the number of common friends of i and j in the observed network $y^{(h-1)}$; a high value of this statistic might turn out to increase the probability that i and j create or sustain a friendship tie with each other in the time interval following t_{h-1}. Note that such lagged network statistics can imply statistical dependence among different dyads only in a restricted form: tie changes in (t_{h-1}, t_h) depend via lagged network statistics on other change

events (potentially on different dyads) only if these happened before or at t_{h-1}. Thus, a model that allows dyadic dependence only via lagged network statistics falls into the class of conditional independence models; it can express mutual dependence among dyads but excludes mutual dependence among individual tie changes.

A detailed example is provided by the two network-evolution scenarios shown in Fig. 5.1. The evolution in the upper row could be explained by a conditional independence model that includes a transitivity effect: the lagged network statistic that encodes for each pair of actors (i, j) the number of directed two-paths from i to j assigns the value one to (A, B) and zero to any other pair. Including this statistic in the specification of the tie probabilities can lead to an increased probability of the newly created tie (A, B).

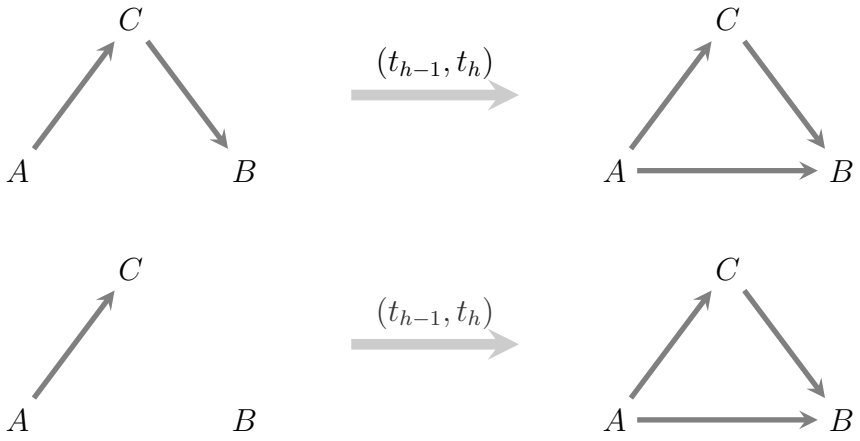


Figure 5.1.: Two different change scenarios in which the networks observed at t_{h-1} (*left*) evolve into the networks observed at t_h (*right*). A conditional independence model could only attribute the evolution in the top row to transitivity; there, the newly created tie (A, B) closes a two-path that was present at the first time point.

The situation is very different in the bottom row of Fig. 5.1. There the same lagged network statistic that counts directed two-paths is zero for all pairs, implying that this statistic does not yield an increased probability of the tie (A, B) — even though (A, B) closes a transitive triangle in the network observed at t_h. In contrast, models that do not rely on conditional independence could interpret the network evolution in the bottom row as a support for transitive closure (assuming that this change pattern has been frequently observed in the data).

5.1.2. Experimental Setup for Network Panel Data

We now illustrate the comparison of conditional independence models with more general models in two established frameworks for modeling network panel data, namely,

temporal exponential random graph models (tERGMs) and stochastic actor-oriented models (SAOMs); refer to Section 2.2 for details on these model frameworks.

To begin with, we make the (Markovian) assumption that the network evolution in the interval (t_{h-1}, t_h) is stochastically determined by the network $y^{(h-1)}$ but conditionally independent of the previous networks $y^{(1)}, \ldots, y^{(h-2)}$. Formally, for all $h \geq 2$ it is

$$P(Y^{(t_h)} = y^{(h)} | y^{(h-1)}) = P(Y^{(t_h)} = y^{(h)} | y^{(1)}, \ldots, y^{(h-1)}) \ . \tag{5.2}$$

This Markovian assumption has nothing to do with the conditional independence assumption that is the focus of our investigation. While conditional independence could also be investigated without the Markov assumption, this would blow up the models and notation.

Temporal exponential random graph models and stochastic actor oriented models, in general, do not fall into the class of conditional independence models. For instance, including a statistic counting triangles in the dependent network implies that incident dyads are mutually dependent.

Therefore, for each ERGM or SAOM statistic that introduces dependency among different ties, we define a lagged statistic that mimics the same structural effect while staying in the class of conditional independence models. Then we compare the quality of models that use only the lagged statistics (and therefore assume conditional independence) with their more general counterparts that use only the original ERGM respectively SAOM statistics.

Formally, for an ERGM statistic s we define the associated lagged network statistic $x(y) \in \mathbb{R}^{n \times n}$ by setting for each pair (i, j)

$$x_{ij}(y) = s(y^{+ij}) - s(y^{-ij})$$

where y^{-ij} denotes the network resulting from y by setting the i, j'th entry to zero and y^{+ij} denotes the network resulting from y by setting this entry to one.

From this lagged network statistic we obtain a lagged ERGM statistic by setting

$$s'(y^{(h)}, y^{(h-1)}) = \sum_{i \neq j} y_{ij}^{(h)} \cdot x_{ij}(y^{(h-1)}) \ , \tag{5.3}$$

i. e., by summing the values of x computed at the preceding time t_{h-1} over all pairs that are connected by a tie at time t_h. We then compare the quality of models that use the ERGM statistic s with those that use the implied lagged statistic s'. For instance, if the ERGM statistic s counts transitive triangles then s and s' assess the evolution in the top row of Fig. 5.1 in an identical way: the newly created tie from A to B closes a transitive triangle in $y^{(h)}$ (which is counted by s) but also closes a directed two path that was present in $y^{(h-1)}$ (which is counted by s'). On the other hand, the evolution in the bottom row is differently assessed by s and s': the ERGM statistic s counts one transitive triangle while the lagged statistic s' assumes a value of zero. A temporal ERGM that includes only statistics of the form given in Eq. (5.3) is a conditional independence model; in this case the model boils down to a binomial

logit model for the occurrence of ties. Indeed, the conditional probability of a tie at time t_h is then independent of the presence or absence of other ties at t_h (albeit it might be dependent on other ties at t_{h-1}).

Analogously, given a SAOM statistic $s_i(y) \in \mathbb{R}$, $1 \leq i \leq n$, the associated lagged network statistic $x(y) \in \mathbb{R}^{n \times n}$ is defined by setting for each pair (i, j)

$$x_{ij}(y) = s_i(y^{+ij}) - s_i(y^{-ij}) \ .$$

This lagged network statistic implies a lagged SAOM statistic by setting

$$s'_i(y, y^{(h-1)}) = \sum_{j=1}^{n} y_{ij} \cdot x_{ij}(y^{(h-1)}), \tag{5.4}$$

i. e., by summing the values of $x(y^{(h-1)})$ over all outgoing ties of actor i in the current network y. We then compare the quality of models that use the SAOM statistic s with those that use the implied lagged statistic s'.

Note that the proposed models will allow parameters $\theta_\ell^{(h)}$ to depend on time step h (i. e., the model is heterogeneous over time). Specifying a time-homogeneous model is straightforward by requiring that $\theta^{(h)} = \theta^{(h')}$ for all pairs of observation points $h, h' \geq 2$.

5.1.3. Design for Interval-Censored Event Data

The conditional independence assumption is not only an issue when dealing with panel data about relational states but also when analyzing interval-censored relational event data (cf. Chapter 2). We note in passing that conditional independence assumptions in models for network event data can be analyzed with quite similar methods. As for panel data we can think of two model frameworks, one based on SAOMs and the other on ERGMs:

In the SAOM-like approach we can compare conditional-independence models (like the one proposed by Brandes et al. [2009c] and Stadtfeld [2012]) with more general models that assume an unobserved but existent order of all events in the same observation interval. It is then allowed that later events depend on previous ones, even if these happen in the same interval. Since the order of events is unobserved a more complicated model estimation has to be applied—similar to the maximum likelihood approach to SAOMs [Snijders et al., 2010a] that augments the data by simulated chains of change events.

In the ERGM-like approach all events in one observation interval are treated as one observed network. Thus, instead of considering the data from one interval as several single-tie events we treat it as one multi-tie event that is modeled by an ERGM-like distribution. This approach would require to dichotomize the so-constructed network event or to extend the ERGM framework to cases where ties are no longer binary variables but encode counting data. Designing such models and comparing them with event network models that assume conditional independence has yet to be done.

5.2. Case Study

In an initial case study, we compare ERG- and SAO-models that assume conditional independence with those that do not, according to the experimental setup that was outlined in Section 5.1.2. As motivated above, it seems possible that the independence assumption becomes less severe when the observation intervals are relatively short compared to the usual rate of change of the analyzed relation. Therefore, the comparison is done with two datasets that differ, among others, in the length of the time interval between observation points: The *Knecht classroom data* describes four friendship networks with inter observation intervals of three months, while the *Newcomb fraternity data* encodes a friendship network observed 15 times, once per week — refer to Section 2.2.1 for details on both data sets.

To allow for straightforward applications of the proposed framework, here, we dichotomize the original Newcomb fraternity data by keeping the top-four nominations for each actor without distinction of their ordering. Note that by this transformation the outdegrees are constant over the actors and over the different time steps. Moreover, regarding the Knecht data, one of the 26 pupils left the class between the second and the third observation point. To obtain networks of constant size, we drop the leaving pupil from all four waves, beacause imputation of ties would be likely to favor those models that are most compatible with the rule of imputation (see Huisman and Steglich [2008]).[1]

Detailed model specifications for the case study are provided in the next section. The assessment criteria then are, firstly, how well these models explain the observed data (Section 5.2.2) and, secondly, whether substantive findings about network effects on the evolution of ties are affected by the (potentially invalid) assumption of conditional independence (Section 5.2.3).

5.2.1. Model Specifications

Model estimation in the ERGM framework is done with the help of the R-package **ergm** [Handcock et al., 2011] which is part of the **statnet** package [Handcock et al., 2003]. For estimating the SAOMs, we use the R-package **RSiena** [Ripley et al., 2012]. In both frameworks the specification is done by choosing the statistics that model the conditional probability of the network $y^{(h)}$, given the previous network $y^{(h-1)}$. We define and estimate three different models that we first describe intuitively. The *basic* model controls the edge density, inertia, and dependence on (actor- and dyad-)attributes. Two models extend the basic model and are compared with each other: The *dyadic-dependence* model introduces reciprocity, indegree, outdegree, and transitivity effects. The *conditional-independence* model mimics the network effects

[1] Note that this level of missingness is very low compared to the levels considered in Huisman and Steglich [2008]. To further assess the impact of dropping one actor we left out others, one by one, to obtain artificially reduced networks of 24 actors. The main result about the log-likelihood differences (Section 5.2.2) was never affected by this data reduction and parameter significance (Section 5.2.3) changed only in rare cases.

of the dyadic-dependence model by substituting each statistic by its conditionally-independent counterpart, as it was described in Section 5.1.2. Our main question is whether the—statistically and computationally much simpler—conditional independence model performs as good as the dyadic-dependence model. Details are provided in the following.

(basic model) ERGMs control the density by the `edges` statistic that counts the number of edges in the network $y^{(h)}$; in SAOMs this is achieved by counting the number of outgoing ties of each actor. The *lagged edges* statistic is included in ERGMs and controls the inertia of ties from one time point to the next. Technically it is the ERGM statistic computed by Eq. (5.3) from the adjacency matrix of the previous wave. Thus, the lagged edges statistic counts the number of edges in $y^{(h)}$ that are also present in $y^{(h-1)}$. A lagged edges statistic is not necessary for SAOMs, since these model inertia implicitly by letting the sequence of tie changes start from network $y^{(h-1)}$.[2] All models contain four attribute-dependent statistics for the Knecht data (none for the Newcomb data): We use the main effects of the "sex" variable on outgoing and incoming ties (controlling whether boys tend to have more in-/out-edges than girls), the homophily effect of the "sex" variable (controlling whether pupils tend to have friends of the same gender), and a covariate term that counts those edges connecting pupils that were in the same primary school.

The dyadic-dependence model adds four statistics to the basic model. The `mutual` statistic counts the reciprocated ties in the network $y^{(h)}$. Furthermore, in the ERGM framework we include the statistics named `gwesp`[3] and `gwidegree` for both datasets and additionally `gwodegree` for the Knecht data.[4] We discuss the choice of the shape parameter [cf. Hunter and Handcock, 2006] below. Within the SAOM framework we include the statistics named `transitive ties` and `indegree popularity` for both datasets and additionally `outdegree activity` for the Knecht data.

The conditional independence model substitutes the four additional statistics of the dyadic-dependence model by their conditionally-independent counterparts computed by the procedure described in Section 5.1.2.

Some technical details All ERGMs were estimated with an MCMC sample size set to 20,000. The conditional independence models could in principle be estimated by ordinary logistic regression; this is not necessary since the estimation method of **ergm** recognizes the independence and automatically switches to a faster estimation procedure. To find appropriate shape parameters for the statistics `gwesp`, `gwidegree`, and `gwodegree` we applied the following steps. We first estimated all models without

[2] Including a lagged edges statistic in SAOMs results often in a non-convergent model.

[3] `gwesp` stands for geometrically weighted edgewise shared partners; cf. Section 2.1.3.

[4] The dichotomized Newcomb data has constant outdegree implying that the *lagged* gwodegree statistic is a constant multiple of the `edges` statistic—leading to a non-identifiable model.

fixing the shape parameters (i. e., by estimating curved exponential random graph models). For many time steps this resulted in poor convergence; for those steps where we did achieve good convergence, the shape parameter for the `gwesp` statistic was typically small but positive (around 0.1) and the estimated shape parameters for the `gwidegree` and `gwodegree` statistics were typically in the range $[0.4, 0.8]$. We chose to fix the shape parameter for `gwesp` at 0.1 and for the other two at 0.5. With this choice all models converged for all time steps.

All actor-oriented models were estimated with the maximum likelihood estimation method proposed in [Snijders et al., 2010a]. For the majority of the models and time steps, convergence was excellent in terms of the criterion (proposed in Ripley et al. [2012]) of *t-ratios for convergence* less than 0.1 in absolute value, and for almost all models, convergence was still satisfying (indicated by t-ratios less than 0.2).

5.2.2. Evaluation of Conditional Likelihood

Since model parameters are estimated via maximum likelihood estimation, a natural way to quantify whether model M explains the evolution from network $y^{(h-1)}$ to $y^{(h)}$ better than model M' is to compare the conditional probabilities of $y^{(h)}$, given $y^{(h-1)}$, i. e.,[5]

$$P_{M;\theta}(Y^{(t_h)} = y^{(h)}|y^{(h-1)}) \quad \text{and} \quad P_{M';\theta'}(Y^{(t_h)} = y^{(h)}|y^{(h-1)}) \ ,$$

where θ and θ' are the maximum likelihood estimates when fitting the models M, respectively M', to the data $(y^{(h)}|y^{(h-1)})$. Since we use the same pair of networks for model estimation and assessment, the conditional likelihoods can only be used to compare models with the same number of parameters.[6] Two models are compared by the difference of the respective log-likelihoods[7]

$$\log(P_{M;\theta}(Y^{(t_h)} = y^{(h)}|y^{(h-1)})) - \log(P_{M';\theta'}(Y^{(t_h)} = y^{(h)}|y^{(h-1)})) \ ,$$

being positive when M is the better model and negative when M' is better.

Concretely, we compare for each dataset and time step the dyadic-dependence model with the conditional independence model as described in Section 5.2.1. The differences in the log-likelihood are summarized graphically in Figure 5.2, using time-dependent bar charts. An upwards pointing black bar indicates that the dyadic-dependence model is better than the associated conditional-independence model; conversely, downwards pointing red bars indicate that the conditional independence model performed better. The height of a bar represents the magnitude of the difference, where the scaling on the vertical axis is such that the maximum difference is normalized with

[5] For readability we denote with "model M" a *class* of ERGMs respectively SAOMs specified by a fixed selection of statistics; such a class of models is specialized to a concrete model by fixing the parameters.

[6] Note that a model with $n(n-1)$ parameters could just remember the value of every tie variable in $y^{(h)}$ and, thus, "predict" the observed outcome with probability one.

[7] Likelihoods of SAOMs are not directly computable. However, the data augmentation approach proposed in Snijders et al. [2010a] allows for estimating differences between log-likelihoods of two models.

respect to data and model framework.[8] Time advances from left to right, and for each estimation step there are two bars: differences in the SAOM-framework (slightly darker shades) are depicted to the right of differences in the ERGM-framework. Raw numbers are provided in Table 5.1.

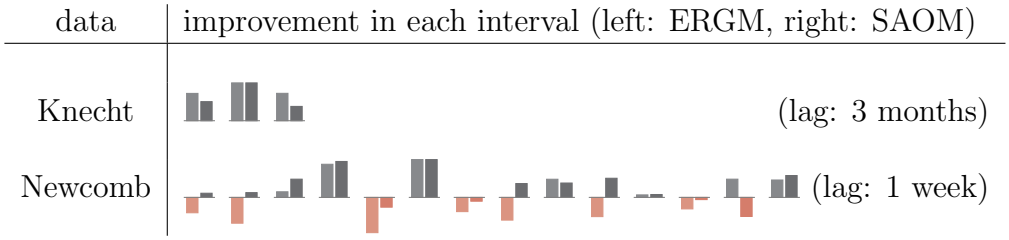


Figure 5.2.: Upward pointing black bars indicate improvement in the log-likelihood when using dyadic-dependence models instead of associated conditional-independence models; downward pointing red bars indicate that the conditional-independence model performs better in the respective time step.

Knecht classroom data (upper part) and Newcomb fraternity data (lower part)

6.37	14.95	6.41											
0.88	2.50	0.60											
-0.83	-1.76	0.25	2.69	-3.00	3.43	-0.75	-1.45	1.04	-1.13	0.10	-0.57	1.05	0.98
0.03	0.04	0.15	0.32	-0.08	0.33	-0.03	0.11	0.12	0.16	0.02	-0.02	-0.16	0.18

Table 5.1.: Log-likelihood differences for both datasets (top rows for ERGMs, bottom rows for SAOMs, respectively). Positive values indicate a better performance of the dyadic-dependence models—indicated by an upward pointing bar in Fig. 5.2.

In summary, dyadic-dependence models perform better than their conditionally-independent counterparts. In the Knecht data, with an inter-observation time of three months, there are no exceptions to this rule. On the other hand, the superiority of dyadic-dependence models diminishes with shorter inter-observation time. In the Newcomb data the dyadic-dependence ERGMs improve over their conditionally-independent counterparts in seven out of 14 time steps; in the SAOM framework the dyadic-dependence models show an advantage in ten out of 14 time steps. By the Markov assumption, the probability of the joint Newcomb data is

$$P(y^{(2)}, \ldots, y^{(15)} | y^{(1)}) = P(y^{(2)}|y^{(1)}) \cdot P(y^{(3)}|y^{(2)}) \cdots P(y^{(15)}|y^{(14)}) \ ,$$

where model parameters might change from interval to interval. Dyadic-dependence models achieve in both frameworks a (slightly) better joint likelihood: the sum of the log-likelihood differences is 0.045 for ERGMs and 1.175 for SAOMs.

[8] We further transformed the absolute differences d by setting the height proportional to $\log(d+1)$ which better reveals the sign of small values.

Newcomb data with a longer time lag How would conditional independence models have performed on the Newcomb data, if it had been collected with the same inter-observation time as the Knecht data? We might approach this question by separately fitting models to the pairs of networks $(y^{(13)}|y^{(1)})$, $(y^{(14)}|y^{(2)})$, and $(y^{(15)}|y^{(3)})$, ignoring the eleven intermediate observations.[9] Models allowing for dyadic-dependence outperformed conditional-independence models on this long-lagged Newcomb data. The log-likelihood differences are 3.93, 15.97, and 5.79 for ERGMs and 0.76, 1.25, and 0.35 in the SAOM framework.[10]

We further varied the preprocessing steps for the Newcomb data by considering only ties that are present in two consecutive waves. That is, we defined a network $y^{(1\&2)}$ by intersecting the first two networks $y^{(1)}$ and $y^{(2)}$ as well as a network $y^{(14\&15)}$ by intersecting the last two networks $y^{(14)}$ and $y^{(15)}$.[11] These networks no longer have constant outdegrees (even though the outdegrees are constrained to be less than or equal to four). The resulting panel data is more comparable to the Knecht data with respect to the inter-observation time and the amount of network change.[12] When fitting models to the data $(y^{(14\&15)}|y^{(1\&2)})$ we found, again, that dyadic-dependence models outperform conditional independence models with a log-likelihood difference of 9.02 for ERGMs and 1.05 for SAOMs. Any variant of the long-lagged Newcomb data that we considered can be better explained by dyadic-dependence models. Results in this section, thus, suggest that the conditional independence assumption becomes more unrealistic with increasing inter-observation time.

5.2.3. Sensitivity of Derived Network Effects

Another aspect on which we compare conditional-independence models with more general models are differences and similarities in derived significant network effects. The question is whether tests of hypothetical network effects are affected by the potentially invalid assumption of conditional independence. To shed light on this question we compare estimated parameters and significance levels for the conditional independence models and associated dyadic-dependence models described in Section 5.2.1.

[9] The time difference between Wave 1 and 13 is 13 weeks (since there is a time gap of two weeks between Wave 9 and 10, due to a holiday break at that time) and, thus, roughly corresponds to three months.

[10] Note that these differences are higher than any difference that results from the one-week steps in the respective model frameworks.

[11] Thus, there is a tie from i to j in $y^{(1\&2)}$ if and only if $y_{ij}^{(1)} = 1$ and $y_{ij}^{(2)} = 1$.

[12] The Jaccard index (in this case measuring the ratio of stable ties) is 0.304, while the mean Jaccard index for the Knecht data is 0.412. For comparison, the mean Jaccard index for the short-lagged Newcomb data is much higher, 0.681.

	DD-ERGM	CI-ERGM	DD-SAOM	CI-SAOM
sex ego	0.581 (0.274) *	0.582 (0.319)	0.454 (0.256)	0.324 (0.267)
	0.421 (0.523)	0.854 (0.286) **	0.618 (0.226) **	0.725 (0.267) **
	0.731 (0.269) **	0.939 (0.291) **	0.477 (0.201) *	0.593 (0.206) **
sex alter	-0.432 (0.325)	-0.344 (0.323)	-0.398 (0.267)	-0.345 (0.274)
	0.081 (0.606)	0.148 (0.285)	0.071 (0.250)	0.238 (0.231)
	-0.101 (0.295)	0.274 (0.290)	0.194 (0.206)	0.351 (0.226)
sex homophily	0.676 (0.263) *	0.898 (0.340) **	0.678 (0.268) *	0.741 (0.286) **
	1.235 (0.566) *	1.258 (0.296) ***	1.082 (0.233) ***	1.000 (0.234) ***
	0.978 (0.253) ***	1.162 (0.298) ***	0.689 (0.204) ***	0.810 (0.212) ***
primary school	-0.218 (0.291)	0.138 (0.360)	0.147 (0.281)	0.138 (0.276)
	0.770 (0.484)	1.070 (0.313) ***	0.568 (0.237) *	0.587 (0.248) *
	0.864 (0.281) **	0.800 (0.319) *	0.492 (0.207) *	0.591 (0.225) **
mutual	1.737 (0.396) ***	1.376 (0.347) ***	1.479 (0.298) ***	1.151 (0.274) ***
	1.396 (0.786)	0.728 (0.306) *	0.882 (0.281) **	0.493 (0.232) *
	1.253 (0.361) ***	0.820 (0.278) **	0.913 (0.198) ***	0.626 (0.193) **
activity ego	-1.898 (0.201) ***	0.445 (0.565)	0.054 (0.023) *	-0.028 (0.026)
	-4.117 (0.243) ***	1.048 (0.513) *	0.102 (0.018) ***	-0.025 (0.023)
	-0.663 (0.306) *	0.689 (0.435)	0.015 (0.016)	-0.025 (0.016)
popularity alter	0.357 (0.912)	-1.765 (0.803) *	0.089 (0.047)	0.221 (0.060) ***
	-1.254 (1.596)	-3.014 (0.965) **	0.043 (0.044)	0.082 (0.041) *
	-0.229 (0.997)	-0.425 (0.940)	0.008 (0.041)	0.065 (0.036)
transitivity	0.569 (0.025) ***	0.333 (0.164) *	0.723 (0.237) **	0.254 (0.187)
	0.679 (0.055) ***	0.193 (0.161)	0.594 (0.353)	0.123 (0.172)
	0.620 (0.028) ***	0.108 (0.222)	0.630 (0.268) *	0.273 (0.198)

Significance levels are marked as follows: *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$.

Table 5.2.: Knecht classroom data. Comparison of estimated parameters of dyadic-dependence (DD) with conditional-independence (CI) models. For each block the parameters and their standard errors are listed from the first time step down to the third.

Knecht data Estimated parameters are shown in Table 5.2. In summary, conditional independence models yield nearly the same results for the attribute-based effects and for the assessment of reciprocity but lead to totally different conclusions for the other (more) structural network effects.

More detailed, all models agree that in all time steps there is evidence for `sex homophily` but no significant parameters associated with the `sex alter` statistic (measuring whether boys receive more ties than girls). The `sex ego` effect (measuring whether boys tend to initiate more ties) is confirmed in the majority of cases and, with one exception, tie probabilities among pupils that have visited the same primary school are significantly increased in the second and third time step but not in the first.

The support for the reciprocity hypothesis (`mutual` statistic) is nearly unchallenged. With one exception, parameters are significantly positive across model frameworks and irrespective of whether models assume conditional independence or not.

The picture completely changes for the other (more) structural network effects. While the dyadic-dependence ERGMs and SAOMs largely agree with each other, the conditional-independence models yield different conclusions. Tests with these models

are in some cases more conservative (less rejection of the null hypothesis) and in other cases more liberal.

More detailed, the dyadic-dependence models yield in the majority of cases strong evidence for transitive closure and dependence on outdegree (`activity ego`), while their conditionally-independent counterparts mostly fail to establish these effects.[13] The disagreement is reversed when considering the `popularity alter` statistic (modeling the influence of indegree on the probabilities of ties). While the dyadic-dependence models consistently found this effect to be insignificant, conditional independence models reject the null hypothesis in the first two time steps.

The comparison so far suggests that conditional independence should not be assumed when testing structural network effects in the Knecht data; instead more general models based, for example, on the ERGM or SAOM frameworks, should be used.

Newcomb data Results for the Newcomb fraternity data (Table 5.3) do not provide such clear distinction as for the Knecht classroom data. There is some disagreement between conditional independence models and their general counterparts, but—in accordance with our expectations—less serious than for the Knecht data. In contrast to what has been found on the Knecht data, there is also considerable disagreement between dyadic-dependence ERGMs and dyadic-dependence SAOMs.

More detailed, the transitivity statistic `gwesp` of the dyadic-dependence ERGMs yields a significant parameter in nine of fourteen time steps—albeit with varying signs. In contrast, the three other models agree with each other by not rejecting the null hypothesis in any time step. There is more agreement with respect to the two other network effects. Parameters associated with the `popularity alter` statistic (dependence on indegree) have been found significant by all models in four time steps and, with some exceptions, mostly insignificant in the other intervals. Finally, the parameters of the `mutual` statistic (measuring reciprocity) are found to be significant by all models in three time steps and mostly insignificant in the others.

Altogether, the occasional disagreement between conditional-independence models and their more general counterparts seems to be less severe in the Newcomb data than in the Knecht data. This provides further support to what we have found in the previous section: with increasing inter-observation times, the gap between dyadic-dependence and conditional independence models becomes more apparent.

[13] Note that *negative* parameters associated with the ERGM statistics `gwodegree` and `gwidegree` point to preferential attachment with respect to outgoing, respectively incoming ties; cf. Hunter [2007].

	DD-ERGM	CI-ERGM	DD-SAOM	CI-SAOM
transitivity	0.278 (0.046) ***	0.160 (0.177)	0.252 (0.287)	0.084 (0.216)
	0.138 (0.048) **	0.157 (0.190)	0.358 (0.315)	0.214 (0.244)
	-0.361 (0.053) ***	0.216 (0.230)	-0.596 (0.349)	-0.083 (0.297)
	0.780 (0.061) ***	0.265 (0.214)	0.645 (0.444)	0.146 (0.315)
	0.654 (0.045) ***	0.382 (0.222)	0.208 (0.335)	0.317 (0.273)
	0.019 (0.052)	-0.028 (0.239)	0.275 (0.379)	0.053 (0.309)
	-0.042 (0.058)	-0.153 (0.270)	-0.681 (0.433)	-0.605 (0.407)
	0.085 (0.059)	0.013 (0.249)	-0.180 (0.437)	0.027 (0.342)
	0.053 (0.059)	0.087 (0.271)	0.410 (0.452)	0.533 (0.389)
	-0.283 (0.063) ***	0.166 (0.254)	-0.728 (0.411)	-0.135 (0.326)
	0.004 (0.047)	0.296 (0.212)	0.141 (0.345)	0.106 (0.305)
	-0.180 (0.072) *	0.170 (0.286)	0.041 (0.405)	0.133 (0.379)
	0.175 (0.040) ***	-0.020 (0.214)	0.382 (0.348)	0.085 (0.268)
	-0.176 (0.050) ***	0.036 (0.209)	0.227 (0.355)	0.250 (0.281)
popularity alter	-0.810 (1.088)	0.124 (0.800)	0.077 (0.073)	0.018 (0.067)
	-1.284 (1.103)	-0.429 (0.932)	0.107 (0.075)	0.099 (0.073)
	-1.257 (1.235)	-0.353 (1.064)	0.301 (0.107) **	0.293 (0.107) **
	-1.626 (0.818) *	-1.443 (1.099)	0.124 (0.083)	0.138 (0.089)
	-1.179 (0.781)	-1.636 (0.960)	0.170 (0.066) **	0.168 (0.063) **
	-3.719 (0.899) ***	-2.630 (0.959) **	0.230 (0.073) **	0.225 (0.079) **
	-0.457 (0.990)	-0.206 (0.803)	0.134 (0.084)	0.124 (0.075)
	-2.232 (0.912) *	-3.135 (1.137) **	0.234 (0.095) *	0.170 (0.087) *
	-1.663 (0.923)	-0.989 (0.951)	0.029 (0.088)	0.007 (0.083)
	0.179 (1.212)	0.467 (0.874)	0.006 (0.096)	-0.025 (0.084)
	-2.522 (0.872) **	-2.845 (1.387) *	0.238 (0.076) **	0.289 (0.088) **
	-1.230 (1.121)	-0.117 (1.028)	0.002 (0.087)	0.011 (0.082)
	-3.544 (0.841) ***	-3.777 (1.267) **	0.166 (0.068) *	0.223 (0.068) **
	-1.744 (0.869) *	-0.966 (0.750)	0.131 (0.073)	0.113 (0.069)
mutual	0.369 (0.440)	0.824 (0.388) *	0.685 (0.344) *	0.762 (0.334) *
	1.962 (0.497) ***	1.368 (0.396) ***	1.248 (0.358) ***	1.249 (0.350) ***
	0.889 (0.514)	0.589 (0.469)	1.072 (0.471) *	0.659 (0.463)
	0.758 (0.544)	0.836 (0.470)	1.147 (0.490) *	0.995 (0.478) *
	0.803 (0.502)	0.602 (0.447)	0.730 (0.388)	0.713 (0.379)
	0.923 (0.550)	0.544 (0.504)	0.805 (0.459)	0.373 (0.448)
	1.112 (0.558) *	0.835 (0.546)	0.554 (0.613)	0.693 (0.611)
	-0.164 (0.652)	0.737 (0.596)	0.515 (0.555)	0.726 (0.536)
	1.398 (0.567) *	1.255 (0.547) *	1.662 (0.575) **	1.348 (0.562) *
	-0.007 (0.601)	0.155 (0.534)	0.053 (0.612)	0.085 (0.556)
	1.205 (0.533) *	0.176 (0.494)	0.960 (0.494)	0.520 (0.472)
	1.035 (0.668)	0.628 (0.581)	0.868 (0.598)	0.889 (0.596)
	1.880 (0.561) ***	1.447 (0.459) **	1.324 (0.432) **	1.483 (0.427) ***
	0.896 (0.522)	0.393 (0.488)	1.015 (0.452) *	0.416 (0.456)

Significance levels are marked as follows: *** for $p < 0.001$, ** for $p < 0.01$, and * for $p < 0.05$.

Table 5.3.: Newcomb fraternity data. Comparison of estimated parameters of conditional-independence (CI) models with dyadic-dependence (DD) models. For each block the parameters and their standard errors are listed from the first time step down to the 14th time step.

5.3. Discussion

The goal of this investigation was to assess the implications of assuming conditional independence of tie-change events when analyzing given longitudinal network data. Based on the quantitative results in our case study, we confidently hypothesize that the data characteristic mostly responsible for the validity or invalidity of the conditional independence assumption is the inter-observation time. In the examples that we considered, three-months between observations yields data that seems to be inappropriate for conditional independence models. The more general dyadic-dependence models consistently achieved a better likelihood for the Knecht data and the artificially thinned Newcomb data. This performance gap diminished in the shorter spaced Newcomb data. Note that both datasets encode network evolution among a group of pupils, respectively students, in the period after these meet for the first time—a situation that is likely to be characterized by relatively fast tie changes. In situations where ties are more stable, longer inter-observation times could still be consistent with the conditional independence assumption. In contrast, relations that change faster could yield network data with considerable intra-observation dependence, even if it is collected at a high frequency.

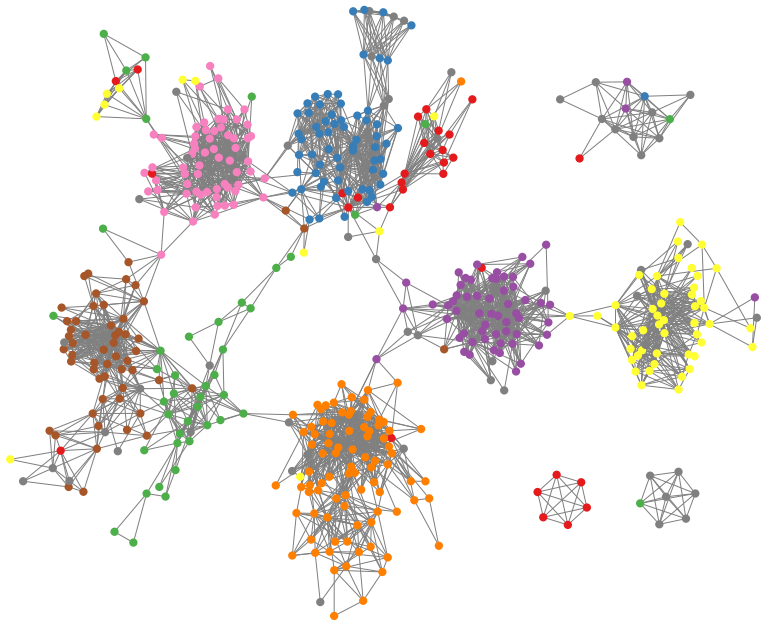
The question whether hypothetical network effects can alternatively be tested with a conditional independence model gets a clear negative answer for the structural network effects (such as transitivity or preferential attachment) in the data with a three-month lag. Structural effects in the Knecht data are quite consistently assessed by the general ERGMs and SAOMs; in contrast, their conditionally independent counterparts lead to very different conclusions. This discrepancy gets somewhat attenuated in the Newcomb data with an inter-observation time of one week.

Tentatively we observe that there was considerable agreement across the two model frameworks, ERGMs and SAOMs, in the Knecht data but less in the Newcomb data. We are not aware of any studies that clarify whether temporal ERGMs and SAOM should confirm or reject the same network effects when applied to the same data. A deeper comparison of ERGMs and SAOMs is certainly a promising task for future work.

Chapter 6.

Exploiting Detailed Timestamps in Relational Data

6



The findings presented in this chapter appeared in Lee et al. [2013].

One important aspect toward a better understanding of evolving social networks is to foresee which connections among actors are likely to be established next. This problem of predicting previously unobserved interactions, based on the network(s) of previously observed interactions, is refered to as *link prediction*.

In their seminal paper which introduced the link prediction problem for social networks, Liben-Nowell and Kleinberg [2003] conducted experiments in large co-authorship networks. In this setting, the exact sequence and spacing of publication dates can hardly be relevant for predicting future interactions because publications dates are distorted anyway (backlogs, preprints, etc.). The aggregation of publication events on a coarser time-scale thus does not appear to be problematic, and might even be necessary to provide accurate predictions.

Due to the problem setting in Liben-Nowell and Kleinberg [2003], however, the practice to convert network event data, if available, into (interval-censored) network panel data has become the predominant approach in the field. That is, today's state-of-the-art link prediction utilizes combinations of complex features derived from network panel data [Lichtenwalter and Chawla, 2012, Lü and Zhou, 2011].

Arguably, whether the conversion from event to panel data is justifiable or not depends crucially on mechanisms which drive tie formation in a given network. In situations different form the one in Liben-Nowell and Kleinberg [2003], available fine-grained temporal information may be highly relevant, making the conversion from event data to panel data difficult to justify because it may destroy important patterns of interaction. For instance, if we are trying to foresee whether node A will send an email message to B in the near future, then an extremely useful piece of information is whether B has *recently* sent A a message; if so, it is likely that A will respond to B soon. This response-mechanism is known as *reciprocity*, and has been observed to be highly relevant for understanding events in social networks [Brandes et al., 2009c]. By aggregating communication events into cross-sectional graphs, traditional link prediction schemes are generally prone to miss such simple and useful mechanisms.

In this chapter, we demonstrate that link predictors can be made more effective and efficient if they operate directly on appropriate time-stamped dyadic event data, and as a result can take advantage of the information contained in the spacing and ordering of relational events. The approach we introduce is based on keeping track of how out of date a node A is with respect to another node B with respect to time-respecting information flow; for doing so we employ the concept of *vector clocks* [Fidge, 1988, Mattern, 1989]. Our results confirm that dyadic features that exploit fine-grained temporal information can be highly relevant for predicting which actors will communicate for the first time in the near future, and are not limited to reciprocity.

In the next section, we describe the types of event data for which we expect fine-grained temporal information to be relevant for link prediction. Then, we review the link prediction problem in this setting, paying particular attention to supervised link prediction, the framework we employ here. We then specify how a modified version of *social* vector clocks [Kossinets et al., 2008] can, among others, be used as a supervised link predictor, and thereby advance the understanding of evolving social networks.

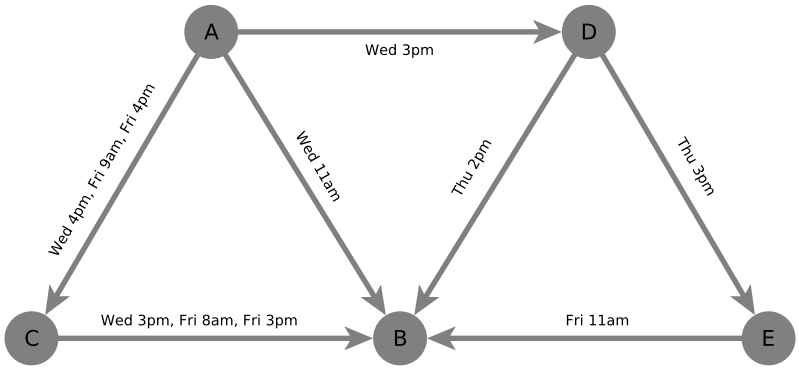


Figure 6.1.: Illustrative example of social network data containing fine-grained information on dyadic *communication* events. Indirect information might flow on time-respecting paths, i.e. along labels that respect the ordering of time. Adapted from Kossinets et al. [2008].

6.1. Information Diffusion via Indirect Updates

We have indicated that the ordering and spacing of communication events might contain valuable additional information over and above the mere number of contacts between a pair of actors. Consider the example in Figure 6.1, which depicts a series of directed communication events, such as e-mail messages. Let us imagine that the actor represented by node A is in charge of organizing a wellness weekend-trip for a group of friends, and that she keeps changing her mind about when and where to go. She finally settles on a plan on Thursday at noon, and all of the subsequent messages she sends out include the final trip details. We can ask: which nodes can possibly know them, given the observed interactions? Clearly, node A communicated with node C after she made up her mind on Thursday, so node C would have received the final information on Friday at 9am. Because node C subsequently sent node B a message, node B could also have received the correct information. On the other hand, nodes D and E could not have received information from node A that is more recent than Wednesday at 3pm.

Two key and related concepts present in this example are *latency* and *indirect updates*. We expect that groups of people who coordinate some action, such as a wellness weekend-trip, will need to synchronize their knowledge of certain key information such as departure time and destination. So in some sense (which we will define formally in Section 6.3) the members of this group have a low latency with respect to each other. Indirect updates, such as the one that made B aware of A's latest trip plans, are an important mechanism for maintaining these low latencies: even though B did not have any direct message from A after she made up her mind, B still got the latest plans indirectly via C.

This type of indirect communication is common in social systems: consider the case of several adult siblings who communicate with each other rarely, and more often communicate with their parents. In this case, the siblings' information on each other can remain up to date due to the central role of the parents, who provide the siblings with an indirect means of communication. More generally, we observe that gossip – the information exchanged when two people talk about an absent third party – is a form of communication prevalent in society and is in essence a form of indirect update.

The motivation behind our approach in this chapter is to demonstrate that information diffusion via indirect updates can be exploited to infer future direct relations. In terms of the example above, we might predict that the siblings are likely to communicate with each other because their latencies with respect to each other remain low. However, the current approach to link prediction throws away much of these temporal clues by first converting any given dyadic event stream into network panel data.

In the datasets we analyze here, we have reason to believe that fine-grained diffusion patterns are relevant to link prediction. For instance, two of the datasets that we will use during our evaluation contain sequences of micro-blogging events that come from Twitter; cf. Section 2.3.1. Bakshy et al. [2011], e.g., have found that word-of-mouth information in Twitter spreads via many small cascades of tweets, mostly triggered by ordinary individuals. These small chains of diffusion are exemplary of the indirect updates we described above, and by considering the details of how information spread, we may be able to infer which nodes will come into direct contact in the future. Detailed information on the datasets we use in our final evaluation is given in Section 6.4.1.

6.2. Link Prediction

In this section, we review the basics of link prediction. In particular, we provide an overview of how machine learning models can be used to combine multiple link predictors – a technique called supervised link prediction.

6.2.1. The Problem and its Evaluation

Along the lines of its original formulation by Liben-Nowell and Kleinberg [2003], we formulate the link prediction problem for dyadic event data as follows:

> Given a sequence of communication events in the form of (time, sender, receiver) tuples, predict which pairs of nodes who had no communication (i.e. are disconnected) in the time interval $[t_0, t_1)$ will communicate (i.e. become connected) in the time interval $[t_1, t_2)$.[1]

[1] In practice, the specification of a link prediction task involves more details, such as whether directed or undirected dyads are considered; we address these points in Section 6.4.2.

An *unsupervised link predictor* is a function which, given a dyad and the list of all previously occurring events, returns a score, where a higher score indicates that a tie is more likely to form in the dyad. *Common neighbors* is an example of an unsupervised link predictor: given a dyad (A,B), return the number of contacts shared by A and B. Although common neighbors is simple, it is quite effective and many of the most effective unsupervised link predictors (such as the similarity measure proposed by Adamic and Adar [2003]) are also based on shared neighbors [Liben-Nowell and Kleinberg, 2003, Lü and Zhou, 2011].

By running a link predictor on all dyads that are disconnected in the interval $[t_0, t_1)$, one can rank all of the possible new links. To evaluate a link predictor, we compare this ranking with the set of new links that actually occur in the period $[t_1, t_2)$. In practice, performance on the link-prediction task is often measured using ROC curves (i.e., true positive rate vs. false positive rate for varying thresholds) or measures based on precision (i.e., true positive rate only). In their recent paper on evaluation in the link prediction problem, however, Lichtenwalter and Chawla [2012] convincingly argue that due to the extreme class imbalance present in the link prediction task, precision-recall curves (including the fraction of detected links) are a more relevant and less deceptive way to measure performance. For that reason, here we exclusively use precision-recall curves for our evaluation.

6.2.2. Supervised Link Prediction

As link prediction is fundamentally a binary classification problem, it is natural to use the powerful binary classification models that have been developed in machine learning [Bishop, 2006]. The primary advantage of this approach is the ability to combine multiple unsupervised link predictors into one joint prediction model. We will now provide a brief overview of how supervised link prediction works. For an in-depth discussion of supervised link prediction see Lichtenwalter et al. [2010].

As is usual in machine learning evaluation, we train and test our classifier on two separate datasets: we must be careful that the classifier is not trained on the same data that is used to evaluate it. For this reason, supervised link prediction requires a train and test framework as depicted in the bottom half of Figure 6.2.

A link prediction classifier is given a set of features related to each disconnected dyad in the period $[t_0, t_1)$, as well as a label which indicates whether the dyad became connected in the period $[t_1, t_2)$. From this information, it learns a model which relates the dyad features to the probability that a previously disconnected dyad becomes connected. To measure the accuracy of a link predictor, we then create a set of test dyad features in the interval $[t'_0, t'_1)$ and use those to score the test labels in the interval $[t'_1, t'_2)$. We measure how accurately the scored dyads predict the set of test labels using the area under the precision-recall curve (AUPR).

We note that the AUPR of a link predictor can fluctuate greatly: as the behavior of users changes, so does the accuracy of the link predictor. In order to better estimate the typical AUPR attained by a link predictor, we can run this procedure many times; we will refer to each run of the procedure outlined in the bottom panel of Figure 6.2

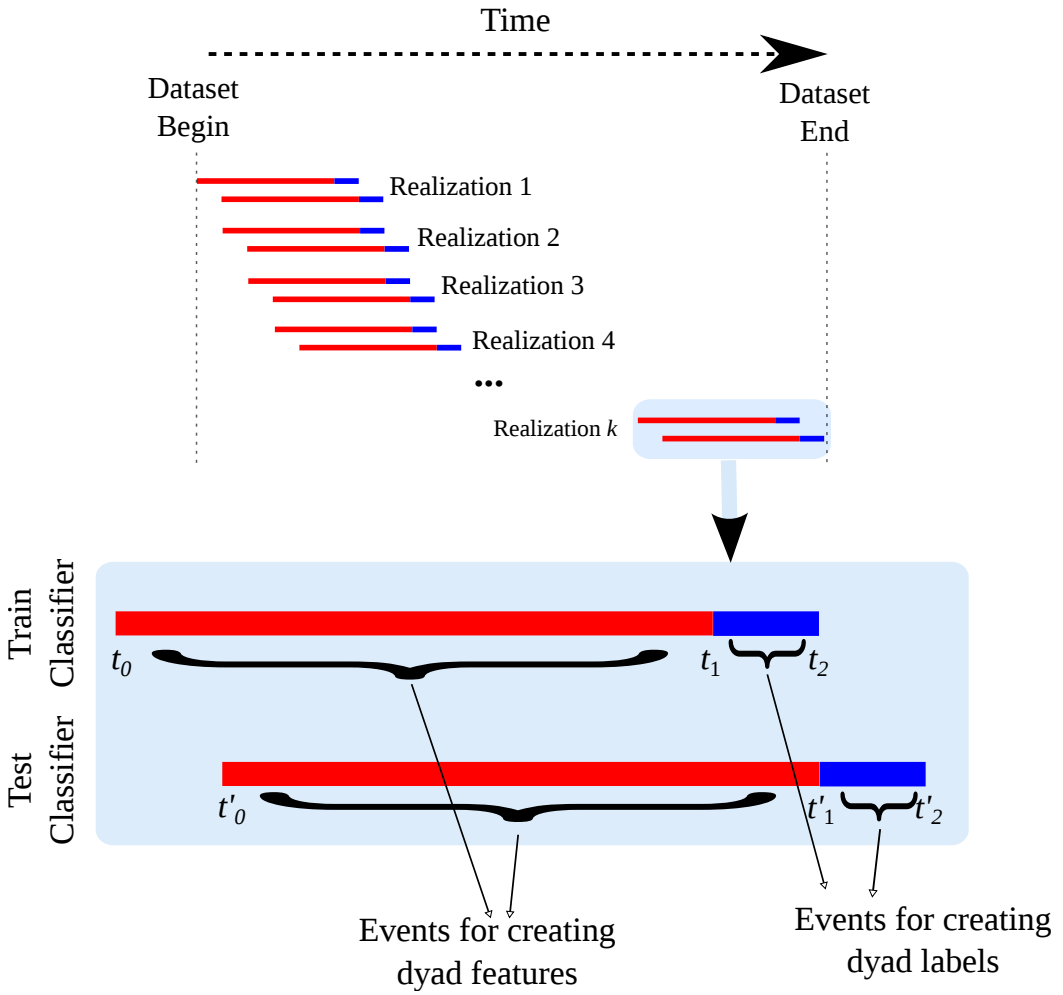


Figure 6.2.: Framework for performing and evaluating supervised link prediction. Each dataset is split into several *realizations*; each realization, in turn, is split into intervals to train and test a classifier.

as a *realization*. As shown in the top panel of Figure 6.2, we shift realizations such that the AUPR of each realization is measured using a distinct set of events.

Lichtenwalter et al. [2010] convincingly argue that the link prediction problem should be stratified over different geodesic distances (i.e., path lengths). That is, the disconnected dyads with a geodesic distance of $N = 2, 3, 4, \ldots$ should each be put into different bins, and a separate classifier should be trained on each bin. This stratification leads to better performance because the decision boundary for each bin can be quite different, with some features (such as common neighbors) being of primary importance at small distances, and other features (such as preferential attachment) becoming more important at larger distances. Thus, by treating each distance as a separate classification task, not only does performance improve, but one can also gain insight into the particular strengths and weaknesses of a predictor. We therefore follow suit and treat each distance as a separate link prediction task.

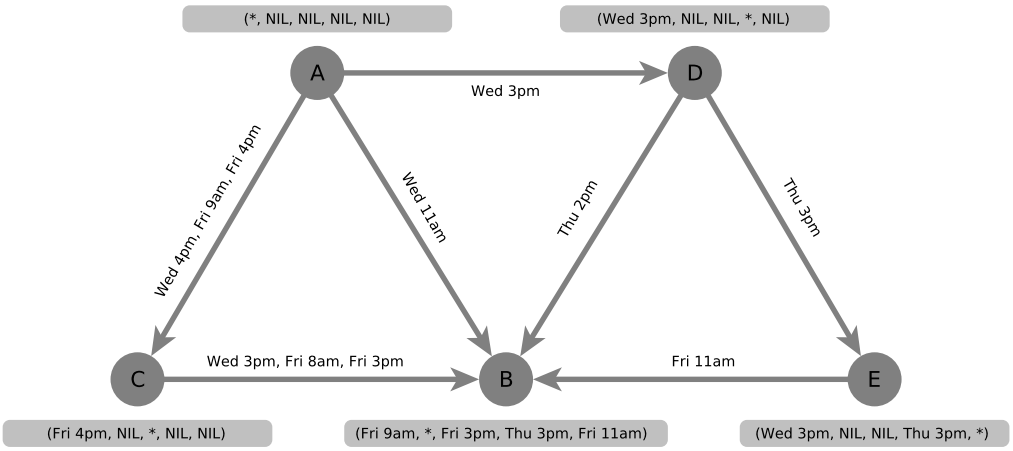


Figure 6.3.: The basic idea of vector clocks: each node's vector clock (in grey) keeps track of the most recent information it could have on the other actors in the network.

6.3. Learning with Vector Clocks

Having introduced the necessary background on link prediction and potential information diffusion via indirect updates, we now explain how fine-grained temporal information can be exploited, using the concept of vector clocks.

6.3.1. Traditional Vector Clocks

Vector clocks were conceptually defined in Fidge [1988] and Mattern [1989] as a means to track causality in concurrent systems, but had implicitly been used before, e.g., in [Parker Jr. et al., 1983], with underlying foundations attributed to Lamport [1978]; for an introduction to vector clock systems in distributed computing refer to [Baldoni and Raynal, 2002]. Kossinets et al. [2008] brought the concept to social network analysis by substituting message-exchanging processes with communicating individuals. In this special setting, the basic motivation of vector clocks is to keep track of the lower bound of how out-of-date a person is with respect to every other person at any time, assuming that information spreads according to a given time-ordered list of communication events.

Reconsider the trip-planning example we introduced in Section 6.1. There we asked: which nodes could possibly know about the most recent details, through either direct or indirect updates? Vector clocks provide a way of answering this question by keeping track of the last possible update that a node could have received from each other node: the vector clock (grey box) next to node E in Figure 6.3 indicates that E cannot possibly have received information from A more recent than Wednesday at 3pm, that it could have received no information whatsoever from nodes B and C, and so on. Each node is always assumed to have up-to date information on itself.

Formally, as already delineated in Section 2.3, let (t_i, s_i, r_i), $i \in \mathbb{N}$, a sequence of

(time,sender,receiver) tuples satisfying $t_i \leq t_{i+1}$ and $s_i \neq r_i$. The set of individuals is defined implicitly by $V = \bigcup_{i \in \mathbb{N}} \{s_i, r_i\}$.

At time t_i, sender s_i and receiver r_i exchange *direct* information about themselves, and *indirect* information about others that result from communication events in the past ($t < t_i$). That is, information can not be forwarded instantaneously but with arbitrary small delay. Now, a *vector clock* is a multivariate function $\phi_{v,t} = (\phi_{v,t}(u) : u \in V)$, in which v's *temporal view* $\phi_{v,t}(u)$ on u at time t is defined as the time-stamp $t^* \leq t$ of the latest information from u that could have possibly reached v (directly or indirectly) until time t. At any time, each actor is up-to-date with respect to itself, $\phi_{v,t}(v) = t$. Temporal views on others can be tracked online as step functions resulting from component-wise maximum calculations of ϕ_{s_i,t_i} and ϕ_{r_i,t_i} at time-steps t_i. Intuitively, ϕ_{s_i,t_i} is updated if and only if a communication event is mutual (such as telephone conversations or meetings), while the update is restricted to ϕ_{r_i,t_i} if a communication event is directed (such as email-, text-, or Twitter-messages).

A major drawback of traditional vector clocks is poor scalability. This results from quadratic space requirements to maintain complete temporal views, along with efficiency problems when performing linear-size maximum calculations on every event. Therefore, traditional vector clocks are too expensive to maintain and manipulate in large graphs. While quadratic space and linear bandwidth requirements are necessary to allow for exact calculations in the general case [Charron-Bost, 1991], approximate calculations of a limited number of temporal views [Torres-Rojas and Ahamad, 1999] and less expansive update algorithms in restricted settings, such as acyclic communication graphs [Meldal et al., 1991], have been proposed. For suitable topologies, additional data structures can be used to reduce the bandwidth of information to be forwarded [Singhal and Kshemkalyani, 1992].

6.3.2. Social Vector Clocks

In contrast to those enhancements stemming from the literature on distributed computing, we propose a modification that is tailored to *social* communication networks. While the original formulation of vector clocks is interesting for social networks because it captures the process of gossip and indirect communication, it does so in an exaggerated and almost clumsy manner. Experiments [Kleinberg, 2000] on the small-world property of social networks [Milgram, 1967, Watts and Strogatz, 1998] and the investigation in Kossinets et al. [2008] suggest that, in the vector clock update algorithm described above, nodes will soon receive huge amounts of information on people they have never met, and whom even their own contacts have never interacted with directly. Indeed, in our own initial experimentation, we found that most actors quickly attain a non-null temporal view with most other actors in the system, and that single communication events often cause an actor to be updated on nearly all of the other actors.

These global updates are hard to justify because they do not seem to resemble social communication. In other words, while exchanging system-wide information is important in the context of distributed computing, such massive information exchanges do

not occur when two people communicate with each other. Rather than updating each other on most of the other actors in the system, the nature of social communication is *bounded* by cognitive limits; such as the number of acquaintances, which does not scale with the size of the overall population [Hill and Dunbar, 2003].

We observe that when two people meet and talk about third parties, they are likely to discuss mutual acquaintances, or at least restrict the conversation to people at least one of them has met directly. Compared to this circle of acquaintances and mutual acquaintances, they are relatively unlikely to talk about any given friend of a friend of a friend, whom neither knows directly. Based on this observation, we propose to bound the reach of indirect updates. Our modification adds one parameter μ to the vector clock framework, which restricts how far information can travel along time-respecting paths; we will also refer to this parameter as the *reach* of indirect updates. More precisely, the reach of indirect information is bounded by the minimal number of hops μ it *ever* took information to travel between a pair of actors on time-respecting paths. Consider the consequence of assigning the following values to μ:

$\mu = 1$ restricts the creation of temporal views to those pairs of actors that have already communicated directly: A node r can receive an indirect update on a node u *if and only if* this *receiver* r has previously had a direct update from u.

$\mu = 2$ additionally allows the creation of temporal views for distance-two neighbors (where distance is measured using time-respecting paths). That is, when considering whether node r can receive an indirect update on a node u via a direct update from node s, e.g., it is always *sufficient* that the *sender* s previously had a direct contact with u. We note that this case has been shown to be especially important in information brokerage [Burt, 2007].

$\mu = \infty$ corresponds to the classical vector clock algorithm with unlimited information spread, and in practice quickly results in quadratic space requirements.

This modification is straightforward to implement because using the vector-clock framework it is trivial to track the length of shortest time-respecting paths: When processing a communication event (t, s, r), the minimum number of hops it ever took information about u to reach r is given by

$$\mathrm{dist}_{t_i}(u, r) = \min(\mathrm{dist}_{t_{i-1}}(u, s) + 1, \mathrm{dist}_{t_{i-1}}(u, r)),$$

where $\mathrm{dist}_t(a, b)$ refers to the length of the shortest time-respecting (a, b)-path until time t. In this way, distances are directed, respecting the ordering of events, and decreasing over the course of time. In our implementation, distances are not known to the source of an information chain, but saved together with the vector clock of the target. Once a short information chain has been observed, a corresponding temporal view is established *and also allowed to be updated by longer information chains.*

Not only does the μ parameter allow the vector clock update process to more closely approximate how indirect updates actually take place in social communication, but in practice this restriction also substantially reduces the memory used by the algorithm,

			$\mu = 1$		$\mu = 2$		$\mu = \infty$	
Dataset	# Events	# Nodes	Mem.	Time	Mem.	Time	Mem.	Time
`irvine`	61734	1899	1	0.9	10	1.4	68	2.5
`election2`	563324	5046	5	4.9	64	20.0	624	55.7
`studivz`	886241	28618	12	3.6	153	11.1	18092	2089
`mobile comm.`	180860015	1238758	1241	1178	17948	23012.4	-	-

Table 6.1.: The runtime and memory requirements of the social vector clocks calculations depends heavily on the reach parameter μ. Memory (in MB) indicates the peak amount of memory needed by our Java implementation; runtime is in seconds.

making it scale to large *sparse* social networks with millions of actors and hundreds of millions of communication events. Table 6.1 shows the memory and CPU runtime requirements of a Java-based implementation of the social vector clocks. The datasets are described below in Section 6.4.1 (except for the proprietary mobile phone communication dataset, which we only use here to demonstrate scalability). These results indicate that setting $\mu = 2$ requires roughly one tenth the memory of traditional vector clocks (where $\mu = \infty$). Further reducing μ to 1 brings the memory requirements down roughly by another factor of 10 — of course, the reduction is not linear and depends on network size and sparsity.

6.3.3. A Link Predictor with Social Vector Clocks

We have described how social vector clocks can, in real time, keep track of the most recent information that could have possibly traveled between pairs of nodes. As one potential application of this framework, we can easily derive several features that may be useful for link prediction. These features can then be combined using a supervised link predictor as we outlined in Section 6.2.2.

A first feature is immediately derived from the temporal views that are saved in the vector clocks: the *current latency* is defined as the difference between the current time and the timestamp saved in the temporal view. As second and third features, we track the number of *direct updates* (i.e., the classical aggregation of relational events) and *indirect updates* that occur between a pair of actors as the vector clocks are computed. A fourth feature we calculate is the *expected latency* between each pair of nodes, which can be thought of as the best guess on how out-of-date an actor is about another at any point in the observation window. See Figure 6.4 for a more detailed illustration of all these features.

Based on our experiences with ranked neighborhood graphs and Simmelian backbones (cf. Chapter 3), however, we do not only keep track of the absolute values of the current and expected latencies, but also keep track of their *ranks from a local perspective*. The rationale is that some users of a service like Twitter may be much more active than others. This heterogeneity will mean that some users will have a latency of weeks with their closest contacts, whereas others will typically have a latency of days or hours with their closest contacts. Such heterogeneity may make it

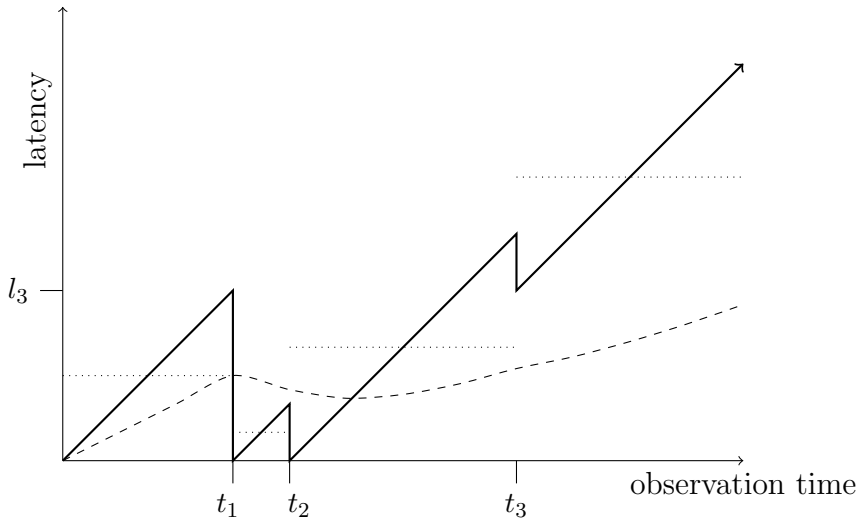


Figure 6.4.: Deriving link prediction features from social vector clocks: Example of a dyad with *two direct updates* (at time t_1 and t_2; the new latency becomes zero), followed by *one indirect update* (at time t_3; the new latency is $l_3 > 0$). The *current latency* (solid line) is a linear jump function resulting from $t - \phi_{v,t}(u)$. The *expected latency* (dashed line) is the weighted mean of average latencies (dotted horizontal lines) between vector clock updates and the current time, respectively.

hard for a classifier to detect a decision boundary. Therefore, from the perspective of a given node i, we sort i's temporal views by their current latency, and then rank the corresponding dyads (cf. Section 3.2.2) — this yields an additional feature for each dyad $\{i, j\}$. We do the same for the expected latency.

So far we have described six features for each dyad: the current latency (both absolute value and rank), the expected latency (absolute value and rank), the number of direct updates, and the number of indirect updates. All of these features can be kept track of in real time and in practice they add little computational overhead. For each dyad, we can keep track of all six of these features in both directions, yielding twelve features. Finally, our definition of social vector clocks included one parameter μ which bounds how far information can travel. In practice one might not know which value of this parameter will lead to the best results; in such a case, one can simply run multiple instances of the vector clocks in parallel, each with a different value of the reach parameter, and combine all the resulting features. A classifier can then learn which feature set is the most informative. For example, in our evaluation below, we run reach-parameterized vector clocks with three different reach parameters: 1, 2, and ∞, which creates a total of 36 features for each directed dyad.

6.3.4. Related Link Prediction Work

While most work on link prediction focuses on the panel-data setting, some previous work also exploits fine-grained temporal information, for example Munasinghe and

Ichise [2011], Sun et al. [2012], and Tylenda et al. [2009].

In Munasinghe and Ichise [2011], the authors propose a new dyadic index that exploits two temporal concepts related to those covered here. In particular, for each dyad (A, B), the index incorporates (1) how much time has elapsed since A and B last interacted, and (2) for each common neighbor C, the similarity of A's and B's latency with C. However, the approach we propose here differs from Munasinghe and Ichise [2011] in a couple of key ways. Firstly, our concept of latency between two nodes allows for indirect updates. Secondly, while in Munasinghe and Ichise [2011] several temporal aspects are combined into a single index, our method keeps these separate and instead produces a feature vector with several dimensions. Given a feature vector with several dimensions, a classifier should have more information to detect the decision boundary.

In Sun et al. [2012], an approach based on meta-paths is introduced that is particularly well-suited for heterogeneous information networks, such as can be formed from co-authorship or product recommendation data. The problem setting in Sun et al. [2012] is not based on *whether* a link will form in the near future, but rather on predicting *when* it will form. While we do not consider this problem formulation here, for certain applications it may be more relevant than the classic link prediction problem; cf. Section 2.3.2.

6.4. Evaluation & Results

6.4.1. Datasets

Two of the datasets we consider come from Twitter. While Twitter is often used as a medium for impersonally broadcasting messages to large numbers of followers, it also supports more targeted forms of communication, in which users explicitly refer to each other. According to the description provided in Section 2.3.1, we filter the two Twitter datasets to include only this targeted form of communication.

Twitter UK Olympics Data The `olympics` dataset covers Twitter communication among a set of 499 UK Olympic athletes over the course of the 18 months leading up to the 2012 Summer Olympic Games, including 730,880 tweets. It was introduced in Greene et al. [2012] and is based on a list of UK athletes curated by *The Telegraph*.[2] We remove all tweets that are not user mentions or retweets between this core set of 499 users, a step which reduces the dataset to 93,613 tweets among 486 users.

Twitter US Elections Data Similar to the `olympics` dataset, the `election` dataset is based on a curated list of Twitter users, in this case curated by Storyful, a commercial news gathering platform targeted at journalists. One of Storyful's features is topical Twitter lists, which journalists can subscribe to in order to

[2] `twitter.com/#!/Telegraph2012/london2012`

		$N = 2$		$N = 3$		$N = 4$	
Dataset	Realizations	Avg. Pos.	Avg. Neg.	Avg. Pos.	Avg. Neg.	Avg. Pos.	Avg. Neg.
`election`	26	796	243658	186	922854	41	1373586
`election2`	26	1664	516415	583	1973244	198	2888931
`olympics`	12	182	15704	38	40257	10	42641
`irvine`	4	478	167674	675	536188	95	348814
`studivz`	23	1332	751172	706	5765692	-	-

Table 6.2.: Link prediction statistics by dataset and path length N. These values report the average over all realizations of each experiment. For each realization, measurements are based on the graph associated with the interval $[t'_0, t'_1)$ and the new edges formed in the interval $[t'_1, t'_2)$.

remain well informed on a given topic. Here we scraped tweets by users on Storyful's US 2012 Presidential election list. In addition we added the Twitter accounts of all those candidates seeking office at the level of governor, US Senator, or member of the US House of Representatives. In total, this dataset covers the date range from Jan. 1st 2012 to Nov. 9th, 2012, and includes 392,662 user mentions and retweets among 2447 Twitter accounts. Additionally, we created an extended version of this dataset, which we will refer to as `election2`, by collecting the tweets associated with Twitter users who were often mentioned in `election`. This extended dataset includes 546,329 tweets among 5,632 users over the same date range as `election`. For more details on this dataset, see Archambault et al. [2013].

StudiVZ Wall posts StudiVZ was created in 2005 as a German competitor for international online social networks. For several years it was the most popular online social network in Germany, although it has eventually been overtaken by Facebook. The `studivz` dataset we examine here is based on a crawl of a single university's subnetwork; it is described in Lee et al. [2011]. While in Lee et al. [2011] the static friendship network is analyzed, here we focus on the wall post ("Pinnwand") data. The dataset is the largest we look at here, containing 26,180 nodes and 886,241 events.

UC Irvine Panzarasa et al. [2009] introduced an event-based dataset which comes from a social networking site set up for the students of the University of California at Irvine. Each event in this dataset is a message—it is unclear whether these are private or public messages. While the dataset covers a period from April to October 2004, the great majority of the events occur between mid April and mid June. Starting in mid June, there is a two week period in which no events occur, and for the remainder of the dataset very few messages are sent. For this reason, we look only at the period from April 10th to June 15th.

6.4.2. Experiment Setup

The High-Performance Link Predictor (HPLP+) introduced in Lichtenwalter et al. [2010] is a state of the art link predictor which combines some of the strongest unsupervised link predictors. In the following experiments, HPLP+ acts as the baseline predictor and our objective is to evaluate the performance of the vector clock link predictor (VCLP) described in Section 6.3.3, and a combined predictor which uses the features from both VCLP and HPLP+.[3]

Framing a supervised link prediction task requires several parameters. One important parameter is the choice of classifier: as in Lichtenwalter et al. [2010], we used bagged forests, a technique suited for the extremely imbalanced classes found in link prediction. However, rather than bagging ten random forests, we bag ten Stochastic Gradient Boosting classifiers. We use the implementation provided in the scikit-learn python package [Pedregosa et al., 2011], using 1000 trees in each classifier, setting the learning rate to 0.005, and subsampling rate to 0.5. In each bag we sampled from the positive instances with replacement, and undersampled from the negative instances with replacement such that the class imbalance ratio was 10 negative for every positive.

Another important experimental parameter is whether the link prediction is directed or undirected. In all of our datasets, edge direction is highly relevant—for example, I might mention President Obama in a tweet, but Obama mentioning me in a tweet would have a completely different meaning. For this reason, we restrict our evaluation to directed link prediction. As one can see in Table 6.2, as the directed geodesic distance N in the link prediction task increases, classes become severely imbalanced, and in the case of `olympics`, hardly any new links form. In `olympics` in general, the classifier has very little positively labeled data to train on, which increases the risk of overfitting when extra features are added.

We must also specify some parameters related to the width of the temporal windows used in the evaluation. In principle, we wanted to make the duration of training period long enough so that a clear and stable snapshot of the network has emerged, and then evaluate on events that occur just after the end of the training period. Therefore, wherever possible we used a training window of 120 days and a test window of 7 days. (In other words, the width of the red bars in Figure 6.2 is 120 days and the width of the blue bars is 7 days.) However, given the short duration of the UC Irvine dataset, we use a shorter training period than in the other datasets, and are not able to run as many realizations of our evaluation; we set the training period to 28 days. Furthermore, the small size of `olympics` meant that in the seven day test period very few new links emerged, leaving the classifier with too little data to train on. Thus, for `olympics` we set the test width to 14 days.

[3] We use the LPMade link prediction framework to compute HPLP+; this is the author's reference implementation [Lichtenwalter and Chawla, 2011].

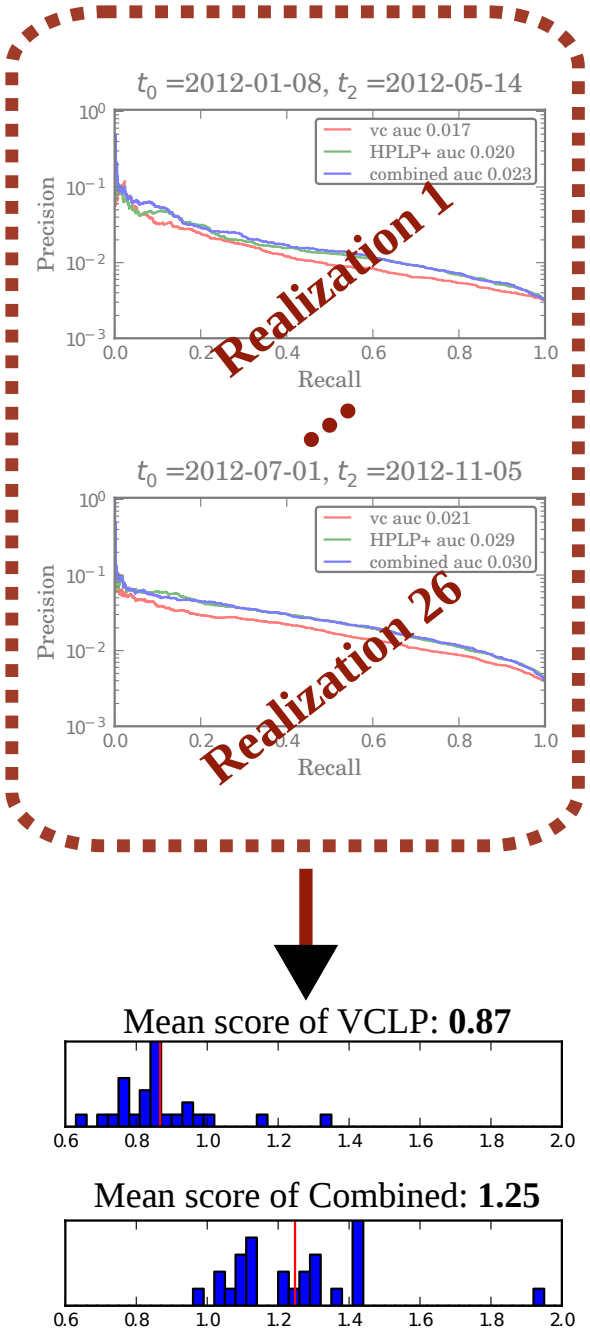


Figure 6.5.: An overview of how we scored the link prediction task. Link predictors are first run on each realization of the experiment. In each realization, precision-recall curves are constructed. The area under the precision-recall curve (AUPR) for each predictor is then measured, and the AUPR of VCLP and the combined link predictor is then divided by the AUPR of HPLP+; this score is the relative performance of each predictor with HPLP+ as the baseline. For each dataset, the average of these scores is then reported; stratified over different geodesic distances.

(a) Results on predicting all edges (at different distances N); see Section 6.4.4 for discussion.

	election			election2			olympics			irvine			studivz	
N	2	3	4	2	3	4	2	3	4	2	3	4	2	3
VC	0.87	0.86	1.05	0.92	0.75	0.86	0.97	1.07	1.06	1.11	1.87	1.12	1.17	1.20
both	1.25	1.15	1.00	1.43	1.38	1.15	1.08	1.10	1.06	1.33	1.65	1.13	1.39	1.22

(b) Results on predicting non-reciprocal edges in the same way; see Section 6.4.5 for discussion.

	election			election2			olympics			irvine			studivz	
N	2	3	4	2	3	4	2	3	4	2	3	4	2	3
VC	0.75	0.78	0.99	0.82	0.65	0.79	0.75	1.02	1.23	0.95	0.79	1.00	0.49	0.57
both	1.17	1.05	1.00	1.38	1.34	1.17	1.06	1.01	1.00	1.19	1.23	1.03	1.11	1.04

Table 6.3.: Average performance of supervised link predictors relative to HPLP+: vector clocks features only (VC), and combination of VCLP and HPLP+ (both).

6.4.3. Experiment Evaluation

The number of realizations performed on each dataset is indicated in Table 6.2. For each realization, we record the precision-recall curve of each predictor, leaving us with a sequence of precision-recall plots such as those presented in the upper section of Figure 6.5. We are interested in how VCLP and the combined predictor perform relative to HPLP+, so we summarize them as follows (as outlined in Figure 6.5): We treat the performance of HPLP+ as the baseline, and in each plot, we measure the area under HPLP+'s precision-recall curve. We then record the area under the precision-recall curve (AUPR) of both VCLP and the combined predictor as a fraction of HPLP+'s AUPR. Thus, if in one realization HPLP+'s AUPR is 0.020 and the combined predictor's AUPR is 0.024, then we record the combined predictor's score as 1.2. After recording this score for all realizations, we are left with a distributions of scores as in the histogram in Figure 6.5. By taking the average of these scores, we can characterize in a single number how much better or worse VCLP and the combined predictor perform than HPLP+. We report these averages for each experiment in Table 6.3; see next two sections for discussion.

6.4.4. Results

In Table 6.2(a), we see that VCLP on its own performs comparably to HPLP+. Considering that HPLP+ combines a broad range of sophisticated graph features, we were surprised to see VCLP perform similarly. Moreover, all network statistics employed by the proposed VCLP can be kept track of online, directly on the list of communication events, while many of the statistics included in HPLP+ have to be recalculated whenever new links are added to the network. Consequently, our results suggest that link prediction with vector-clock statistics can be performed much more

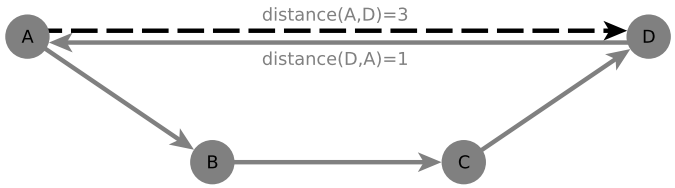


Figure 6.6.: The directed dyad (A, D) has a geodesic distance of 3, but the distance of dyad (D, A) is 1. Thus, the dyad (A, D) would be included in the experiments whose results are presented in Table 6.2(a), but would be excluded from the experiments whose results are presented in Table 6.2(b).

efficiently in any situation, where the model parameters are learned beforehand and applied in real-time on a growing sequence of "test" events.

When the features of VCLP and HPLP+ are combined, the performance increase over HPLP+ is substantial. In general, the performance gain is largest when we are predicting links on dyads that have a geodesic distance $N = 2$. Performance gain decreases for greater N, suggesting that VCLP features are most useful for predicting local links rather than long-range links (`irvine` is an exception to this trend, where $N = 3$ sees by far the largest boost to performance). The improvement is smallest on `olympics`, perhaps because the classifier struggles with the small number of positive training examples.

Given that stand-alone VCLP and HPLP+ yield similar prediction accuracy, it is interesting to observe the added value of combining both predictors. In other words, there appear to be qualitative differences in the network effects that can be captured in the VCLP and HPLP+ framework.

6.4.5. Controlling for Reciprocity

Imagine we're trying to predict whether a node A will soon send its first message to D. One of the features included in VCLP is D's current latency with A through direct updates – in other words, how many seconds have elapsed since D sent a message to A. Given the significance of reciprocity, this feature will be extremely useful for cases where D has just sent a message to A. It could be the case that this feature alone – which is trivial to keep track of without vector clocks – is responsible for all of the benefit that comes from VCLP. In that case, we could simply keep track of this single feature and forget about vector clocks.

To measure whether this is the case, we run the entire evaluation again, but exclude all dyads where D has had any direct contact with A; see Figure 6.6. The results presented in Table 6.2(b) are in the same units as the results presented in Table 6.2(a). We observe that the performance of VCLP does indeed drop, but that there is still a significant benefit provided by combining the features of VCLP and HPLP+. Again, we stress that the lackluster performance on `olympics` may be due to the small number of new links that form, which provides very few positive training examples.

6.5. Discussion

The current approach used by state of the art link predictors is to operate in a panel data setting, in which finer-grained temporal information is ignored. In cases where link formation is not driven by cascades of information, such an approach might be appropriate. Regarding co-authorship networks, for instance, precise information on the sequence and spacing of events may be largely irrelevant or even misleading, and so it may be reasonable to aggregate away information on exact publication dates. However, in some networks – such as the Twitter retweet/mention networks mentioned here – information cascades are an important mechanism for driving the formation of ties. In such a setting, the information contained in the exact sequence and spacing of events is highly relevant. By aggregating all events into a static graph, yet, traditional link prediction schemes cannot even exploit simple and useful mechanisms such as (temporal) reciprocity.

Our results suggest that dyadic features that exploit fine-grained temporal information beyond reciprocity are highly relevant for predicting which actors will communicate for the first time in the near future. The basic idea is to exploit information on how out of date a node A is with respect to another node B, and for doing so we adopt the concept of vector clocks. As an essential modification, we parameterized the traditional vector-clock concept to bound the reach of indirect information. Not only does this make the vector-clock update process more closely approximate how indirect updates actually take place in social communication, but in practice this restriction also dramatically reduces the memory used by the algorithm, thus making it applicable to large *sparse* social networks with millions of actors and billions of communication events.

We have demonstrated that binary classifiers can indeed exploit actor latencies to improve accuracy in link prediction. Even HPLP+, a classifier which utilizes a wide range of graph features based on aggregated panel data, can perform substantially better when provided with additional features based on vector clocks. Moreover, the Vector Clock Link Predictor (VCLP) on its own already performs comparably to HPLP+, which allows for much more efficient link prediction in any situation where the model parameters are learned beforehand and applied in real-time on a growing sequence of events.

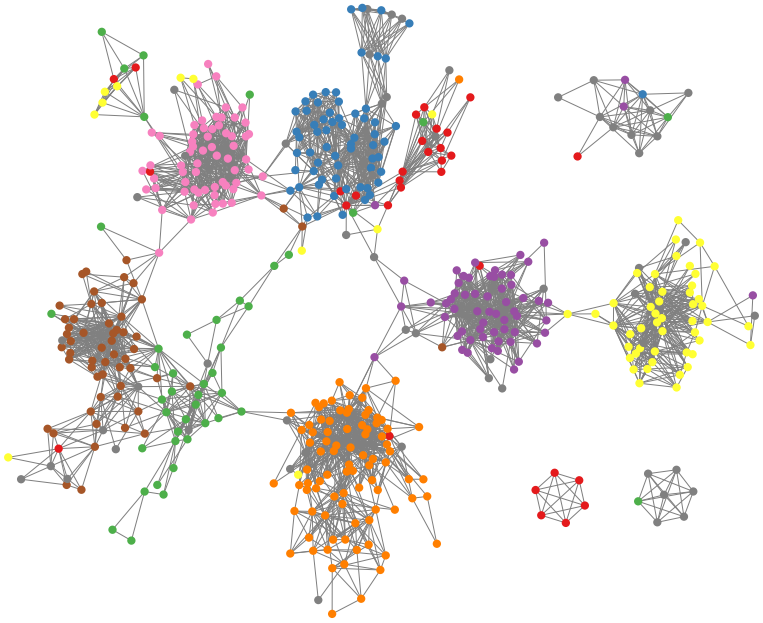
While the intuitive motivation for social vector clocks and their demonstrated performance is already satisfying, we are still keen to gain more detailed insight into the actual mechanisms behind link formation, e.g., by combining feature-selection schemes and even more elaborate substantive theories.

Finally, the application to link prediction has only demonstrated one possible scenario for exploiting the parameterized vector clock update process. We see a wide-open space for many more utilizations, e.g., in the statistical analysis of dyadic event data; cf. Section 2.3.2. In particular, social vector clocks provide a means to transform dyadic event data into tie strength proxies that align well with the extraction of Simmelian backbones (cf. Chapter 3).

Chapter 7.

Conclusion

7



Toward a better understanding of evolving social networks, we emphasized the importance and consequences of varying time granularity in empirical data and presented four methodological contributions:

Simmelian backbones and a corresponding notion of triadic cohesion have been demonstrated to reveal essential structure in (online) social networks. The approach is rooted in substantive sociological theories and its main ingredients — local rankings and subsequent similarity calculations — are intuitive, efficient, and flexible. The presented application has been tailored to community detection, but similar instantiations of ranked similarity calculations might prove valuable in other disciplines such as graph drawing as well. Moreover, the technique carries over to any time granularity in empirical data and thus provides a general tool toward a better understanding of evolving social networks.

Gestaltmatrix visualizations make available an all new technique for representing asymmetric longitudinal network data. Evolution is mapped on space, rather than time, by means of minimalistic combinations of sparklines, multivariate glyphs and Gestalt theory. The approach is thus very different from observable trends toward ever more complex animated, interactive and three-dimensional representations of consecutive network states. The proposed seesaw design is a graphic representation for "ups and downs" in social relationships and meant to leverage visual perception capabilities for the identification and understanding of latent patterns therein. Holistic views of entire dyadic evolutions within matrix cells clearly alleviate the difficulty of retaining a mental map.

Lagged network statistics were incorporated into a general framework that allows to quantify the implications of assuming conditional independence of tie-change events in evolving social networks. The presented case study on network panel data has justified complex dependency assumptions within the stochastic actor-oriented and temporal exponential random graph models, in general, but also revealed important consequences of varying time granularity: the gap between dyadic-dependence and conditional independence models became more apparent with increasing inter-observation times. Over and above, conditional independence models yield nearly the same results for attribute-based effects and the assessment of (dyadic) reciprocity but lead to totally different conclusions for (more) structural network effects such as transitivity.

Social Vector Clocks exploit detailed time information in dyadic event sequences. The proposed parameterization allows the vector clock update process to more closely approximate how indirect updates actually take place in social communication. Moreover, it allows the efficient calculation of information latencies among actors in large *sparse* social networks with millions of actors and hundreds of millions of communication events. A preliminary application to link prediction has demonstrated the added value of correspondingly derived measures, and we see a wide-open space for many more instantiations, e.g., in the statistical analysis of dyadic event data.

With these novel approaches toward a better understanding of evolving social networks, we hope to initiate further substantive applications in the social science. As a common feature of the presented methods, we have stressed that evolving social networks are 'more than the sum' of its constituent dyadic relationships; just as dyadic relationships are 'more than the sum' of individual interactions.

Holistic approaches toward a better understanding of evolving social networks, in general, are in line with the importance of Gestalt theory in the lineage of social network analysis [Scott, 2000, page 8, Figure 2.1]. The here presented methods, in particular, are based on holistic interpretations of *ties, triads, and time*: 1) gestaltlines are meant to communicate entire time information in dyadic tie sequences as a whole, 2) Simmelian backbones and social vector clocks exploit triadic patterns (and available time information) beyond the aggregate level, and, 3) broadly speaking, conditional independence models correspond to summing individual tie-changes while more general dyadic-dependence models provide a holistic interpretation of network evolution over time.

In contrast to these holistic approaches, a discrete (atomistic and static) treatment of network ties is prone to ignore crucial information embedded in structure and time; and thus brings into question the extensive use of dyadic aggregate data across actors, ties, and time. Examples of those include the standard scenario to transform dyadic event data into time-sliced graphs, as well as purely density-based community detection. Indeed, the here presented case-studies — namely, hidden homophily effects in online social networks (cf. Figure 3.7), gestaltlines casting doubt on statements from the literature (cf. Figure 4.8), inappropriate conditional independence assumptions (cf. Figure 5.2), and increased performance in link prediction with social vector clocks (cf. Table 6.3) — revealed deficiencies in previous interpretations of empirical data that we suspect resulted from more aggregate data views.

A problem related to unconditional, dyadic, and static aggregation are fairly strong homogeneity assumptions that current methods rely on. Arguably, however, discriminating features of exceptional actors, dyads, and periods are of great importance in identifying factors that govern the network evolution. In this direction, gestaltmatrix representations allow to "see" longitudinal network data in its very own gestalt and absolute heterogeneity, while ranked neighborhood graphs allow for asymmetry and heterogeneity even in undirected networks with homogeneous relationships.

In summary, a better understanding of evolving social networks is closely linked with the exploitation of structural and temporal patterns beyond the aggregate level analysis of ties — and algorithmic approaches should integrate with long-standing theories and substantive arguments. Simmelian backbones, in particular, acknowledge the fundamental importance of triadic configurations in networks representing social ties and indicate the added value of sociologically informed computational analyses.

Future Work. At the end of the individual chapters, we already discussed challenges and open problems that are specific to the presented approaches. Here, we only repeat that a deeper comparison of temporal exponential random graph models and stochastic actor oriented models is certainly a promising task for future work. In particular, we are not aware of any studies that clarify to which degree the two model frameworks should confirm or reject the same network effects when applied to the same data.

Besides the challenges and consequences arising from network dynamics (i.e., temporal information), network *size* defines a second important dimension that goes along with ever more excessively available data. Yet, as others have noted, "the sophisticated use of stochastic dynamical system models for network formation, particularly in contexts with networks of non-trivial size, is still arguably in its infancy" [Kolaczyk, 2009, p. 193]

At another point, we have touched upon the broad literature on community detection and link prediction. Combining both disciplines seems to be a very interesting direction for future investigations. In particular, Lichtenwalter et al. [2010] convincingly argue that the link prediction problem should be stratified since there might be very different decision boundaries for different types of dyads. Yet, the proposed bins for different geodesic distances (i.e., path lengths) is only one possible instantiation. We expect even better results as soon as dyads are stratified according to previously detected intra- and inter- community dyads; preferably based on a triadic notion of cohesion.

In the light of the aforesaid, furthermore, we note that it is often hard to distinguish different triadic settings in network visualizations of non-trivial size, since both sociomatrix and sociogram rather communicate a notion of density due to the gestalt law of proximity. A workaround is the use of graphical attributes (such as color) to distinguish graphical elements (such as lines) w.r.t. derived triadic weights. Alternative visualizations that allow for more holistic interpretations of triadic evolution – comparable to the seesaw design for dyadic time series – provide another interesting direction for future research.

More general, recently released journals such as *Network Science* [Brandes et al., 2013c] and the *Journal of Complex Networks* [Estrada, 2013] expressly underline that network analysis is an interdisciplinary field expected to have significant impact on a number of disciplines beyond the social science. Bridging the theory gap [Granovetter, 1979] between the face value of network analysis and its methodological stringency, however, remains a particularly challenging endeavor for future research.

Finally, we would like to add a word on gestaltline representations beyond social network data. We feel that the detailed case study presented in the appendix, and its internal validity test show that, as a design principle, the concept of gestaltlines is viable. We see a wide-open space for creative research into glyph design and alignment based on this principle.

Appendix A.

Case Study on Gestaltlines

Our introduction of gestaltline visualizations in Section 4.1, has been accompanied with several examples from the literature. These examples were well-behaved in the sense that a relatively simple data set exhibited relatively clear patterns from known categories.

Here, in the appendix, we turn to another (non-network) data set that is more complex in the number of dimensions and relations between them, and for which no prior knowledge about the presence of a pattern existed. In fact, these data motivated the invention of gestaltlines in the beginning. Our goal is to demonstrate how the design ideas of gestaltlines can be applied in less controlled, and thus more realistic, situations. Moreover, we provide expert feedback on the improved interpretation of their data, and a user study that entails formal evidence on the usefulness and validity of gestaltlines. This work has been published in Brandes et al. [2013b].

A.1. Background and Data

The data originate from a psychological experiment that was conducted as part of a larger study on the influence of early-life stress on human reward processing and decision making. Further details are reported in Steffen [2010].

Thirty subjects participated in the experiment. Eighteen of them were psychiatric patients who experienced early life stress (10 female, average age 39.1 $\pm$ 12.6), half-and-half to a low and high degree. The other twelve subjects form a healthy control group (7 female, average age 43.4 $\pm$ 17.2).

In a repeated measurement design, each participant was subjected to a sequence of 240 computerized gambling trials while being measured for brain activity. As within-subject factors, 10 (euro) cents or 50 cents were at stake in each trial and the announced chance of winning was 10%, 50%, or 90%. Each of the six variants was played 40 times, with the entire sequence in individually randomized order. Subjects had to decide whether they wanted to pass or play. Hence, there are four possible outcomes for each trial: passed, played and won, played and lost, or no decision until a timeout of 2 sec. Each outcome was presented to the subject before the next trial.

With a theoretical maximum gain or loss of $120 \cdot 50\,\text{cents} + 120 \cdot 10\,\text{cents} = 72\,\text{EUR}$, purely rational decisions (play at 90% chance of winning, pass at 10%, and any strategy

at 50%) yield an expected gain of $40 \cdot [0.9 \cdot (50\,\text{cents} + 10\,\text{cents}) + 0.1 \cdot (-50\,\text{cents} - 10\,\text{cents})] = 19.2\,\text{EUR}$. Subjects were given a starting budget of 10 EUR and received an actual payment between 0 EUR and 20 EUR determined from 20 randomly chosen trial outcomes.

The experiment thus generated 30 sequences of 240 four-dimensional data points (stake, chance, decision, outcome) with 24 possible values. Analysis on the aggregate level confirmed some expected differences between patients and control subjects in cortical activation patterns and also in the number of irrational decisions (play at 10% chance of winning, pass at 90%).

Interestingly, however, five of the seven subjects beating the expectation were patients, three of them even with a high stress level. A closer look at the detailed data is supposed to test for systematic effects such as strategy learning or the onset of boredom that do not show on the aggregate level.

A.2. Gestaltline Design

Similar to the examples from Section 4.1.2, we display each subject within a separate gestaltline and represent all 240 four-dimensional data points in sequence order on the horizontal axis left-to-right, i.e. time progresses in Western text reading direction.

To support the domain experts in noticing systematic effects in decision making, we choose glyphs such that different gambling strategies induce holistic forms. Given the noise produced by randomization of game settings, we present the data with regard to a baseline (refer to the example of streaks in sports results in Section 4.1.2), instead of ranking data dimensions by importance and mapping them separately to an equal number of the most accurately perceived graphical features in the same order (cf. Figure 4.1).

There are two data dimensions of utmost importance. On the one hand, the deliberate decision to play, pass, or wait until timeout, is the only data dimension that is non-randomized and directly attributed to the subject. On the other hand, the announced chance of winning is the pivotal determinant of the randomized gambling design, because this information alone determines the sign of the expected gain and should hence figure prominently in any decision. Consequently, we propose a baseline – rationality – that combines these two data dimensions.

Clearly, rationality of a subject's decision is defined relative to announced chance and subsequent decision – rational decisions: play at 90%, pass at 10%; irrational decisions: pass at 90%, play at 10%. However, rationality can not be assessed for decisions at 50% chance of winning, since the expected gain/loss is zero. Instead, we can score such trials as active (play at 50%) or passive (pass at 50%) decisions. Both concepts, rationality and joy of playing, overlap in a sense that extremely active (play at 10%) or passive (pass at 90%) decisions become irrational. Thus, we can order decisions from one extreme (irrational-passive) to the other (irrational-active) – pass at 90%, pass at 50%, pass at 10%, play at 90%, play at 50%, play at 10% – relative to a baseline (rational decisions; pass at 10%, play at 90%).

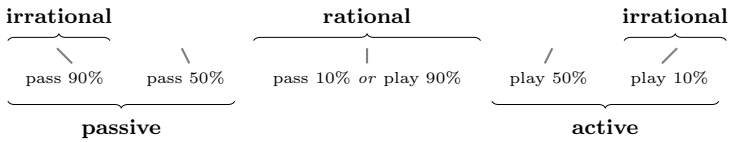


Figure A.1.: **Visualization** and **interpretation** of participant decisions relative to a rational baseline.

A corresponding graphical feature to map this ordering can be borrowed from Sedgewick's illustrations (Section 4.1.1). That is, each trial is represented by a line segment, the slope of which corresponds to the interpretation of that decision with regard to our baseline; cf. Figure A.1. We found this mapping intuitive, since the leaning of line segments suggests a correspondence to the subjects' postures during the experiment, such as sitting straight during rational decisions and leaning forward while gambling.

Integrating the remaining data attributes into these basic glyphs is straightforward. Note that we use the same vertical line segment for both rational decisions, since we do not interpret them as being active or passive. To distinguish these decisions, we exploit that outcome of pass at 10% (no change in a player's budget) is different from play at 90% (change in budget). Concretely, outcomes are represented only by small colored dots (blue for profit, red for loss) in the center rather than coloring the entire line segment to promote strategy information over outcome. For the same reason, irrational decisions are slightly highlighted with a stronger glyph hue.

Moreover, line width and dot size are determined by the amount of money at stake and the occurrence of a timeout is distinguished from authentic pass decisions by the use of white dots. In this way the colored dot areas correspond directly to the change of budget. Using less important graphical features for less important experimental conditions and special outcomes facilitates macro-micro reading [Tufte, 1990] in which the participant strategy is represented as the visually dominant information, and details about the individual trials can be scrutinized if desired.

To increase the effects of continuity and proximity we decided against a monospace arrangement. That is, ||/|//\|/||/|/||||||/|//\\|/||||\/||\||/|/|\///|||/||/|//|\///\/\||/|/|/|//|/|||||/|/|//|/ the (horizontal) space of a single glyph is proportional to the deviations in our baseline (Figure A.1), and we use kerning to reduce space in between consecutive glyphs.

These design decisions take advantage of several gestalt laws: Local trends in decision making go along with similar leaned line segments /// (law of similarity), and little whitespace in between (law of proximity). Still, slight variations in the leaning of line segments resulting from the randomized experiment do not hamper the detection of stable decision patterns /|///|||/||/ (law of continuity). Moreover, strategy changes are promoted by the law of symmetry //\\, and irrational decisions are highlighted with prominent line segments |/| (law of figure and ground). Finally, placing small dots in the center of line segments does not influence the perception of leanings \\ (law of closure).

A.3. Expert Feedback

The visualization of the data as it was presented to the domain experts is reproduced in Figure A.2. They were excited to "*see*" their data in its entirety and commented several times on the gestaltlines' elegance and compactness. In the following we report on some of their interpretations.

The first thing to note is that participants with a net gain around the expected 19.2 EUR played rationally for the most part, which was to be expected. Secondly, there are some obvious differences in strategy. Participant C_1, for instance, employs an aggressive, risk-taking strategy start to finish; whereas subject H_3 starts out aggressively active and acts increasingly errant toward the end, where many irrational decisions go together with the occurrence of a timeout — possibly indicating the onset of boredom. As another example, player L_8 employs a risk-avoiding strategy that occasionally results in irrational decisions. A rare case of learning appears to be participant L_2 for whom the chart calms considerably from the first quarter of games to last quarter of games.

Seeing the entire data in this way, the domain experts found no indication that the analysis on the aggregate level was distorted by unlucky patients or superior learning of control subjects. This was reassuring for the earlier analysis. On the other hand, the domain experts confidently hypothesized that there is more within-class variation among patients than among controls, and that this variation is larger with respect to strategy than performance. While this suggests further substantive work beyond the scope of this study, it should be noted that the strategy hypothesis would have been difficult to arrive at analytically.

A.4. User Study

Feedback from the domain experts supported our driving hypothesis that sparklines made from gestalt theory-informed glyphs facilitate holistic recognition of patterns, trends, and outliers in multivariate sequences. However, the gestaltlines have been designed to resonate with their understanding of the substance. To gather evidence on the general validity of our design, we tested whether it appeals to a broader audience. Concretely, we conducted an unannounced study with a group of 52 second-year undergraduate students in computer science who had no prior background in information visualization.

Subjects were handed a questionnaire with 12 items on the gestaltlines from Section A.2 together with a two-sided A4 handout explaining their design and its basic interpretation. Both handout and questionnaire had been refined following several pretests. The complete user study was limited to 35 minutes, and the fastest subjects needed 17 minutes to read the handout and answer the questionnaire.

Figure A.2.: Sequence data for all 30 subjects with 240 trials each using the gestaltlines as defined in Section A.2. Subjects are partitioned into groups and ordered by net gain/loss (expectation for a rational strategy is 19.2). Two general observations are larger gains for rational subjects and few apparent strategy changes over the course of the experiment. More detailed interpretation in the main text.

	active	passive	rational	irrational
/	92.3%			92.3%
/	94.2%		21.2%	1.9%
|			98.1%	
\		92.3%	21.2%	1.9%
\		90.4%		94.2%

Figure A.3.: Results for introductory question of user study (Section A.4); **Q1** "Please mark for each row the appropriate fields". Unexpected answers highlighted in red.

Initially we tested the subjects' understanding of decisions being represented relative to a baseline defined by expected outcome (Figure A.1). Subjects were asked to classify single decisions (**Q1**) as well as small sequences of glyphs (**Q2**). As summarized in Figure A.3, the vast majority of subjects understood the basic design. In particular, not a single decision was unexpectedly classified as passive or active, and there was only one outlier concerning rationality. We were mildly surprised, however, to find that more than one fifth of the subjects interpret every decision at 50% chance as being rational. On the other hand, subjects were able to match small sequences of rational  (100%), passive  (98.1%), active  (98.1%), and irrational  (96.2%) strategies with the expected label.

Subsequently, similar results were obtained for classifying a complex sequence of rational-passive decisions with noise (**Q4**; 88.5% rational, 78.9% passive), and identifying most (ir)rational strategies among various gestaltlines (**Q7**; 96.2% expected answers). Likewise, subjects were able to find the most prominent strategy change across two gestaltlines (**Q8**; only 3 subjects selected unexpected gestaltline; 7 subjects selected unexpected breakpoint). About 80% of the subjects were able to mark the longest uninterrupted subsequence of rational decisions by player H_2 (**Q9**; Figure A.2).

In **Q3**, we asked subjects to mark all (irrational) outliers within a complex sequence of otherwise rational decisions. Notably, each outlier instance has been marked by at least 86.5% (and up to 94.2%) of the subjects. Somewhat disappointingly one third of the subjects also marked a pass-at-50%-chance decision that was surrounded by play-at-50%-chance decisions  . While this constellation indeed results in perceiving increased leanings of glyphs [Talbot et al., 2012], well-trained readers would have noticed that irrational decisions are slightly highlighted with a stronger glyph hue. Still, subjects found an interesting subsequence (after losing two 50% games, the next one is passed) and potentially have interpreted this temporal change in decision making to be irrationally inconsistent.

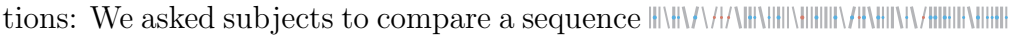
Since an occasional confusion of rationality and consistency was already observed in our pretests, additional questions served to disentangle observations from interpretations: We asked subjects to compare a sequence  of rational decisions with two modified versions of itself. First, all 50% games were redefined as being played  . 88.2% of subjects (expectedly) evaluated the

Figure A.4.: Results for **Q11** "Please divide the sequence into sub-sequences and label them with a brief description" (irrational, rational, **outliers**, passive, active). Answers split into irrational vs. rational (*left*) and passive vs. active (*right*) with **outliers** highlighted. Top two rows indicate our predictions, others the actual answers (one row per subject); representation implemented by T. Döbele and V. Mühlberg.

modified sequence to be more consistent (**Q10**), but 51.9% of subjects also (unexpectedly) evaluated this sequence to be more rational (**Q5**; 38.5% equally rational). Second, the games of the modified gestaltline were permuted to match the length of the original gestaltline. As a result, the majority (54.9%) of subjects evaluated both gestaltlines to be equally rational (**Q12**); that is, only about half of the subjects did realize that the length of gestaltlines is influenced by rationality of decisions, consistency of decisions, and permutation of games alike.

Another impression from the pretests was reproduced in **Q6**: We asked "Did the following subject go lucky?" with regard to a sequence of games that were primarily won, but did involve unfavorable outcomes alike. While the correct answer was "No" (almost all 50% games were lost), about half of the answers were wrong ("Yes"), with explanations such as "more blue dots" / "big wins". That is, subjects did observe correctly the relevant information that needs to be scrutinized (colored dots), but about half of their interpretations did not succeed in disentangling (expected) profit from (unexpected) luck.

Finally, subjects were asked to annotate a short sequence of complex patterns in decision making with strategies, breaks, and outliers (**Q11**). The reassuring results are summarized as gestaltlines themselves in Figure A.4.

The user study thus demonstrates that a short briefing is sufficient for untrained readers to reliably find holistic patterns, outliers and breaks within the proposed visualizations.

A.5. Discussion

Gestaltlines are a conceptual design approach aiming for compact graphical representations of complex data sequences. It consists of the arrangement of especially designed glyphs in sparklines such that patterns of interest yield gestalts which can be perceived holistically and pre-attentively.

The presented case study, and its internal validity test are meant to demonstrate that, as a design principle, the concept of gestaltlines is viable and opens a wide space for creative research. But two examples for exciting future research topics are the

search for alternatives to existing representations such as separation plots [Greenhill et al., 2011] and the extension to hierarchical patterns.

Given the vastness of the potential design space and usage scenarios, however, comprehensive design guidelines are far beyond the scope of the present investigation. Much detailed research will be needed to assess with confidence the effectiveness of gestaltlines for specific data patterns in specific applications. Note that studies such as Fuchs et al. [2013] are likely to yield different results when considering gestalt-informed glyphs and arrangements. It will also be important to understand the relative reading accuracy and efficiency of variant gestaltlines and alternative graphical designs. In addition to task and design-related factors, comparative studies will have to take contextual factors such as the available media into account.

Like every form of visualization, gestaltline design is limited by constraints such as resolution, number of discernible colors, or shape complexity. Where these boundaries are is yet another question that we cannot answer today.

Bibliography

Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211–230, 2003.

Daniel Archambault, Helen Purchase, and Bruno Pinaud. Animation, small multiples, and the effect of mental map preservation in dynamic graphs. *IEEE Transactions on Visualization and Computer Graphics*, 17(4):539–552, 2011.

Daniel Archambault, Derek Greene, and Pádraig Cunningham. Twittercrowds: Techniques for exploring topic and sentiment in microblogging data. *Under Review*, 2013.

Adelchi Azzalini and Adrian W. Bowman. A look at some data on the Old Faithful Geyser. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 39 (3):357–365, 1990.

Eytan Bakshy, Jake M. Hofman, Winter A. Mason, and Duncan J. Watts. Everyone's an influencer: quantifying influence on Twitter. In *Proc. ACM WSDM*, pages 65–74, 2011.

Roberto Baldoni and Michel Raynal. Fundamentals of distributed computing: A practical tour of vector clock systems. *IEEE Distributed Systems Online*, 3(2), 2002.

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

John A. Barnes. Class and committees in a norwegian island parish. *Human Relations*, 7(1):39–58, 1954.

Giuseppe Di Battista, Peter Eades, Roberto Tamassia, and Ioannis G. Tollis. *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall, 1999.

Skye Bender-deMoll and Daniel A. McFarland. The art and science of dynamic network visualization. *Journal of Social Structure*, 7, 2005.

Julian Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society B*, 36(2):192–236, 1974.

Carla Binucci, Ulrik Brandes, Giuseppe Di Battista, Walter Didimo, Marco Gaertler, Pietro Palladino, Maurizio Patrignani, Antonios Symvonis, and Katharina Zweig.

Drawing trees in a streaming model. *Information Processing Letters*, 112(11):418–422, 2012.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, 2008.

Béla Bollobás. *Random Graphs*. Cambridge University Press, 2001.

Stephen P. Borgatti and Daniel S. Halgin. On network theory. *Organization Science*, 22(5):1168–1181, 2011.

Stephen P. Borgatti, Ajay Mehra, Daniel J. Brass, and Giuseppe Labianca. Network analysis in the social sciences. *Science*, 323(5916):892–895, 2009.

Ulrik Brandes and Thomas Erlebach, editors. *Network Analysis: Methodological Foundations*, volume 3418 of *Lecture Notes in Computer Science*. Springer, 2005.

Ulrik Brandes and Bobo Nick. Asymmetric relations in longitudinal social networks. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2283–2290, 2011.

Ulrik Brandes, Jörg Raab, and Dorothea Wagner. Exploratory network visualization: Simultaneous display of actor status and connections. *Journal of Social Structure*, 2(4), 2001.

Ulrik Brandes, Patrick Kenis, and Dorothea Wagner. Communicating centrality in policy network drawings. *IEEE Transactions on Visualization and Computer Graphics*, 9(2):241–253, 2003.

Ulrik Brandes, Patrick Kenis, and Jörg Raab. Explanation through network visualization. *Methodology*, 2(1):16–23, 2006. Spanish translation in *REDES* 9(6), 2005.

Ulrik Brandes, Martin Hoefer, and Bobo Nick. Network creation games with disconnected equilibria. In *Proc. 4th Intl. Workshop on Internet and Network Economics*, volume 5385 of *Lecture Notes in Computer Science*, pages 394–401. Springer, 2008.

Ulrik Brandes, Natalie Indlekofer, Jürgen Lerner, Martin Mader, and Bobo Nick. Panel or event data? Presentation at the 6th Conf. on Applications of Social Network Analysis, unpublished, 2009a.

Ulrik Brandes, Jürgen Lerner, Uwe Nagel, and Bobo Nick. Structural trends in network ensembles. In *Proc. 1st Intl. Workshop on Complex Networks*, volume 207 of *Studies in Computational Intelligence*, pages 83–97. Springer, 2009b.

Ulrik Brandes, Jürgen Lerner, and Tom A.B. Snijders. Networks evolving step by step: Statistical analysis of dyadic event data. In *Proc. of the 2009 Intl. Conf. on Advances in Social Network Analysis and Mining*, pages 200–205, 2009c.

Ulrik Brandes, Jürgen Lerner, Bobo Nick, and Steffen Rendle. Network effects on interest rates in online social lending. In *Proc. INFORMATIK 2011*, volume P-192. GI Edition - Lecture Notes in Informatics (LNI), 2011.

Ulrik Brandes, Natalie Indlekofer, and Martin Mader. Visualization methods for longitudinal social networks and stochastic actor-oriented modeling. *Social Networks*, 34(3):291–308, 2012a.

Ulrik Brandes, Sven Kosub, and Bobo Nick. Was messen Zentralitätsindizes? In M. Hennig and C. Stegbauer, editors, *Die Integration von Theorie und Methode in der Netzwerkforschung*. VS Verlag Sozialwissenschaften, 2012b.

Ulrik Brandes, Linton C. Freeman, and Dorothea Wagner. Social networks. In Roberto Tamassia, editor, *Handbook of Graph Drawing and Visualization*, pages 803–837. CRC Press, 2013a.

Ulrik Brandes, Bobo Nick, Brigitte Rockstroh, and Astrid Steffen. Gestaltlines. *Computer Graphics Forum*, 32(3):171–180, 2013b.

Ulrik Brandes, Garry Robins, Ann McCranie, and Stanley Wasserman. What is network science? *Network Science*, 1(1):1–15, 2013c.

Ronald S. Burt. *Structural holes: The social structure of competition*. Harvard University Press, 1992.

Ronald S. Burt. Structural holes versus network closure as social capital. In *Social Capital: Theory and Research*, pages 31–56. De Gruyter, 2001.

Ronald S. Burt. Secondhand brokerage: Evidence on the importance of local structure for managers, bankers, and analysts. *Academy of Management*, 50(1):119–148, 2007.

Carter T. Butts. A relational event framework for social action. *Sociological Methodology*, 38(1):155–200, 2008.

John M. Chambers, William S. Cleveland, Beat Kleiner, and Paul A. Tukey. *Graphical Methods for Data Analysis*. Chapman and Hall, 1983.

Dempsey Chang, Laurence Dooley, and Juhani E. Tuovinen. Gestalt theory in visual screen design: a new look at an old subject. In *Proceedings of the 7th World Conference on Computers in Education*, CRPIT '02, pages 5–12, 2002.

Bernadette Charron-Bost. Concerning the size of logical clocks in distributed systems. *Inf. Process. Lett.*, 39(1):11–16, 1991.

Norishige Chiba and Takao Nishizeki. Arboricity and subgraph listing algorithms. *SIAM Journal on Computing*, 14(1):210–223, 1985.

Nicholas A. Christakis and James H. Fowler. The spread of obesity in a large social network over 32 years. *The New England Journal of Medicine*, 357(4):370–379, 2007.

Mei C. Chuah and Stephen G. Eick. Glyphs for software visualization. In *Proc. of the 5th Intl. Workshop on Program Comprehension*, pages 183–191, 1997.

William S. Cleveland. *Visualizing Data*. Hobart Press, 1993.

William S. Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

Ethan Cohen-Cole and Jason M. Fletcher. Detecting implausible social network effects in acne, height, and headaches: longitudinal analysis. *BMJ*, 337(a2533), 2008a.

Ethan Cohen-Cole and Jason M. Fletcher. Is obesity contagious? Social networks vs. environmental factors in the obesity epidemic. *Journal of Health Economics*, 27(5): 1382–1387, 2008b.

Carlos D. Correa and Kwan-Liu Ma. Visualizing social networks. In Charu C. Aggarwal, editor, *Social Network Data Analytics*, pages 307–326. Springer, 2011.

Skyler J. Cranmer and Bruce A. Desmarais. Inferential Network Analysis with Exponential Random Graph Models. *Political Analysis*, 19(1):66–86, 2011.

Mark J.C. Crescenzi. Reputation and interstate conflict. *American Journal of Political Science*, 51(2):382–396, 2007.

Wouter de Nooy. Networks of action and events over time. a multilevel discrete-time event history model for longitudinal network data. *Social Networks*, 33(1):31–40, 2011.

David Dekker. Measures of simmelian tie strength, simmelian brokerage, and the simmelianly brokered. *Journal of Social Structure*, 7(1), 2006.

Josep Díaz, Jordi Petit, and Maria Serna. A survey of graph layout problems. *ACM Computing Surveys*, 34(3):313–356, 2002.

Patrick Doreian, Roman Kapuscinski, David Krackhardt, and Janusz Szczypula. A brief history of balance through time. *Journal of Mathematical Sociology*, 21:113–131, 1996.

Patrick Doreian, Vladimir Batagelj, and Anuška Ferligoj. *Generalized Blockmodeling*. Cambridge University Press, 2005.

Christopher DuBois and Padhraic Smyth. Modeling relational events via latent classes. In *Proc. 16th ACM SIGKDD intl conf on Knowledge Discovery and Data mining*, pages 803–812, 2010.

Peter Eades, Wei Lai, Kazuo Misue, and Kozo Sugiyama. Preserving the mental map of a diagram. In *Proc. of Compugraphics*, pages 24–33, 1991.

Niklas Elmqvist, Thanh-Nghi Do, Howard Goodell, Nathalie Henry, and Jean-Daniel Fekete. ZAME: Interactive large-scale graph visualization. In *Proc. IEEE Pacific Visualization Symposium (PacificVis 2008)*, pages 215–222. Springer, 2008.

David Eppstein and Emma S. Spiro. The h-index of a graph and its application to dynamic subgraph statistics. In Frank Dehne, Marina Gavrilova, Jörg-Rüdiger Sack, and Csaba D. Tóth, editors, *Algorithms and Data Structures*, volume 5664 of *Lecture Notes in Computer Science*, pages 278–289. Springer, 2009.

Paul Erdős and Alfréd Rényi. On random graphs. *Publicationes Mathematicae Debrecen*, 6:290–297, 1959.

Margarita Esponda-Argüero. Techniques for visualizing data structures in algorithmic animations. *Information Visualization*, 9(1):31–46, 2010.

Ernesto Estrada. Journal of complex networks: Quo vadis? *Journal of Complex Networks*, 1(1):1–2, 2013.

Alex Fabrikant, Ankur Luthra, Elitza Maneva, Christos H. Papadimitriou, and Scott Shenker. On a network creation game. In *Proc 22nd annual symposium on Principles of Distributed Computing*, pages 347–351. ACM, 2003.

Ronald Fagin, Ravi Kumar, and D. Sivakumar. Comparing top k lists. In *Proc of the 14th annual ACM-SIAM Symposium on Discrete Algorithms*, pages 28–36, 2003.

Elena Fanea, Sheelagh Carpendale, and Tobias Isenberg. An interactive 3D integration of parallel coordinates and star glyphs. In *Proceedings of the 2005 IEEE Symposium on Information Visualization*, pages 149–156, 2005.

Stephen Few. *Information Dashboard Design : The Effective Visual Communication of Data*. O'Reilly, 2006.

Colin J. Fidge. Timestamps in message-passing systems that preserve the partial ordering. In *Proc. 11th Australian Computer Science Conference*, pages 56–66, 1988.

Karl Flieder and Felix Mödritscher. Foundations of a pattern language based on gestalt principles. In *Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, pages 773–778, 2006.

Elaine Forsyth and Leo Katz. A matrix approach to the analysis of sociometric data: preliminary report. *Sociometry*, 9:340–347, 1946.

Santo Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

Santo Fortunato and Marc Barthelemy. Resolution limit in community detection. *PNAS*, 104(1):36–41, 2007.

A. Frank and A. Asuncion. UCI machine learning repository, 2010. URL `http://archive.ics.uci.edu/ml`.

Ove Frank and David Strauss. Markov graphs. *Journal of the American Statistical Association*, 81(395):832–842, 1986.

Linton C Freeman. The sociological concept of 'group': An empirical test of two models. *American Journal of Sociology*, pages 152–166, 1992.

Linton C. Freeman. Visualizing social networks. *Journal of Social Structure*, 1(1), 2000.

Noah E. Friedkin. *A Structural Theory of Social Influence*. Cambridge University Press, 1998.

Johannes Fuchs, Fabian Fischer, Florian Mansmann, Enrico Bertini, and Petra Isenberg. Evaluation of alternative glyph designs for time series data in a small multiple setting. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, 2013.

Mohammad Ghoniem, Narendra Jussien, and Jean-Daniel Fekete. VISEXP: visualizing constraint solver dynamics using explanations. In *FLAIRS'04: 17th intl Florida Artificial Intelligence Research Society conf*, pages 263–268. AAAI press, 2004.

Mohammad Ghoniem, Jean-Daniel Fekete, and Philippe Castagliola. On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. *Information Visualization*, 4:114–135, 2005.

Edgar N. Gilbert. Random graphs. *Annals of Mathematical Statistics*, 30(4):1141–1144, 1959.

Debra S. Goldberg and Frederick P. Roth. Assessing experimentally derived interactions in a small world. *PNAS*, 100(8):4372–4376, 2003.

Sanjeev Goyal. *Connections: an introduction to the economics of networks*. Princeton University Press, 2009.

Mark S. Granovetter. The strength of weak ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

Mark S. Granovetter. The theory-gap in social network analysis. In P. Holland and S. Leinhardt, editors, *Perspectives on Social Network Research*, pages 501–518. Academic Press, 1979.

Derek Greene, Derek O'Callaghan, and Pádraig Cunningham. Identifying topical Twitter communities via user list aggregation. *arXiv preprint arXiv:1207.0017*, 2012.

Brian D. Greenhill, Michael J. Ward, and Audrey E. Sacks. The separation plot: A new visual method for evaluating the fit of binary models. *American Journal of Political Science*, 55(4):991–1002, 2011.

Thomas L. Griffiths and Zoubin Ghahramani. The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224, 2011.

Mark S. Handcock. Assessing degeneracy in statistical models of social networks. *Center for Statistics and the Social Sciences, University of Washington*, Working Paper No 39, 2003.

Mark S. Handcock and Krista J. Gile. Modeling social networks from sampled data. *Annals of Applied Statistics*, 4(1):5–25, 2010.

Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, and Martina Morris. *statnet: Software Tools for the Statistical Modeling of Network Data*, 2003. URL `http://statnetproject.org`. Version 2.0.

Mark S. Handcock, David R. Hunter, Carter T. Butts, Steven M. Goodreau, Pavel N. Krivitsky, and Martina Morris. *ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks*, 2011. URL `http://statnetproject.org`. Version 2.3.

Steve Hanneke, Wenjie Fu, and Eric P. Xing. Discrete temporal models of social networks. *Electronic Journal of Statistics*, 4:585–605, 2010.

Martin Harrigan. Using vector clocks to visualize communication flow. In *Proc of the Intl Conf on Advances in Social Networks Analysis and Mining*, pages 241–247, 2010.

Mark Harrower and Cynthia A. Brewer. ColorBrewer.org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–30, 2003.

Freda-Marie Hartung and Britta Renner. Perceived and actual social discrimination: The case of overweight and social inclusion. *Frontiers in Psychology*, 4(147):1–5, 2013.

Christopher G. Healey, Kellogg S. Booth, and James T. Enns. Visualizing real-time multivariate data using preattentive processing. *ACM Transactions on Modeling and Computer Simulation*, 5(3):190–221, 1995.

Jeffrey Heer and Maneesh Agrawala. Multi-scale banking to 45 degrees. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):701–708, 2006.

Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 203–212, 2010.

Jeffrey Heer, Nicholas Kong, and Maneesh Agrawala. Sizing the horizon: The effects of chart size and layering on the graphical perception of time series visualizations. In *Proc SIGCHI Conf on Human Factors in Computing Systems*, pages 1303–1312, 2009.

Fritz Heider. Attitudes and cognitive organization. *Journal of Psychology*, 21:107–112, 1946.

Marina Hennig, Ulrik Brandes, Jürgen Pfeffer, and Ines Mergel. *Studying Social Networks: A Guide to Empirical Research*. Campus Verlag, 2012.

Nathalie Henry, Jean-Daniel Fekete, and Michael J. McGuffin. NodeTrix: A hybrid visualization of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1302–1309, 2007.

Russell A. Hill and Robin I.M. Dunbar. Social network size in humans. *Human Nature*, 14(1):53–72, 2003.

Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, 2012.

Danny Holten and Jarke J. van Wijk. A user study on visualizing directed edges in graphs. In *Proceedings of the 27th international conference on Human factors in computing systems*, CHI '09, pages 2299–2308, 2009.

Werner Horn, Christian Popow, and Lukas Unterasinger. Support for fast comprehension of ICU data: visualization using metaphor graphics. *Methods of Information in Medicine*, 40(5):421–424, 2001.

Mark Huisman. Imputation of missing network data: Some simple procedures. *Frontiers in Psychology*, 10(1), 2009.

Mark Huisman and Christian E.G. Steglich. Treatment of non-response in longitudinal network studies. *Social Networks*, 30:297–308, 2008.

David R. Hunter. Curved exponential family models for social networks. *Social Networks*, 29:216–230, 2007.

David R. Hunter and Mark S. Handcock. Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*, 15(3): 565–583, 2006.

David R. Hunter, Steven M. Goodreau, and Mark S. Handcock. Goodness of fit for social network models. *Journal of the American Statistical Association*, 103: 248–258, 2008.

Alexander Ihler, Jon Hutchins, and Padhraic Smyth. Adaptive event detection with time-varying poisson processes. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 207–216, 2006.

Matthew O. Jackson. *Social and Economic Networks*. Princeton University Press, 2008.

Brian Karrer and Mark E.J. Newman. Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, 83(1):016107, 2011.

Michael Kaufmann and Dorothea Wagner, editors. *Drawing Graphs: Methods and Models*. Springer, 2001.

Daniel A. Keim. Designing pixel-oriented visualization techniques: Theory and applications. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):59–78, 2000.

Jon Kleinberg. The small-world phenomenon: an algorithm perspective. In *Proc. ACM SToC*, pages 163–170, 2000.

Bryan Klimt and Yiming Yang. Introducing the enron corpus. In *Proc. 1st Conference on Email and Anti-Spam*, 2004.

Andrea Knecht. *Friendship Selection and Friends' Influence. Dynamics of Networks and Actor Attributes in Early Adolescence*. PhD thesis, University of Utrecht, 2008.

Eric D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer Series in Statistics. Springer, 1st edition, 2009.

Robert Kosara and Silvia Miksch. Metaphors of movement: a visualization and user interface for time-oriented, skeletal plans. *Artificial Intelligence in Medicine*, 22(2): 111–131, 2001.

Robert Kosara, Christopher G. Healey, Victoria Interrante, David H. Laidlaw, and Colin Ware. Thoughts on user studies: Why, how, and when. *IEEE Computer Graphics and Applications*, 23:20–25, 2003.

Gueorgi Kossinets. Effects of missing data in social networks. *Social Networks*, 28(3): 247–268, 2006.

Gueorgi Kossinets, Jon Kleinberg, and Duncan Watts. The structure of information pathways in a social communication network. In *Proc. of the 14th ACM SIGKDD intl conf on Knowledge Discovery and Data mining*, pages 435–443, 2008.

David Krackhardt. Simmelian Ties: Super Strong and Sticky. In Roderick Kramer and Margaret Neale, editors, *Power and Influence in Organizations*, pages 21–38. Sage, 1998.

David Krackhardt. The ties that torture: Simmelian tie analysis in organizations. *Research in the Sociology of Organizations*, 16:183–210, 1999.

David Krackhardt and Mark S. Handcock. Heider vs simmel: Emergent features in dynamic structures. In Edoardo Airoldi, David M. Blei, Stephen E. Fienberg, Anna Goldenberg, Eric P. Xing, and Alice X. Zheng, editors, *Statistical Network Analysis: Models, Issues, and New Directions*, volume 4503 of *Lecture Notes in Computer Science*, pages 14–27. Springer, 2007.

Leslie Lamport. Time, clocks, and the ordering of events in a distributed system. *Commun. ACM*, 21(7):558–565, 1978.

Paul F. Lazarsfeld and Robert K. Merton. Friendship as a social process: A substantive and methodological analysis. In Morroe Berger, Theodore Abel, and Charles H. Page, editors, *Freedom and Control in Modern Society*, pages 18–66. Van Nostrand, 1954.

Conrad Lee and Pádraig Cunningham. Benchmarking community detection methods on social media data. *arXiv preprint arXiv:1302.0739*, 2013.

Conrad Lee, Thomas Scherngell, and Michael J. Barber. Investigating an online social network using spatial interaction models. *Social Networks*, 33(2):129–133, 2011.

Conrad Lee, Bobo Nick, Ulrik Brandes, and Pádraig Cunningham. Link prediction with social vector clocks. *Proc. 19th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 782–792, 2013.

Jürgen Lerner, Natalie Indlekofer, Bobo Nick, and Ulrik Brandes. Conditional independence in dynamic networks. *Journal of Mathematical Psychology*, 57(6):275–283, 2013.

Kevin Lewis, Marco Gonzalez, and Jason Kaufman. Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*, 109 (1):68–72, 2012.

David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proc. of the 12th Intl. Conf. on Information and Knowledge Management*, pages 556–559. ACM, 2003.

Ryan N. Lichtenwalter and Nitesh V. Chawla. LPmade: Link Prediction Made Easy. *Journal of Machine Learning Research*, 12:2489–2492, 2011.

Ryan N. Lichtenwalter and Nitesh V. Chawla. Link prediction: fair and effective evaluation. In *Proc. ACM/IEEE ASONAM*, pages 376–383, 2012.

Ryan N. Lichtenwalter, Jake T. Lussier, and Nitesh V. Chawla. New perspectives and methods in link prediction. In *Proc. ACM CIKM*, pages 243–252, 2010.

Alessandro Lomi and Francesca Pallotti. Relational collaboration among spatial multipoint competitors. *Social Networks*, 34(1):101–111, 2012.

Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.

Dean Lusher, Johan Koskinen, and Garry Robins. *Exponential Random Graph Models for Social Networks: Theory, Methods, and Applications*. Structural Analysis in the Social Sciences. Cambridge University Press, 2013.

Jock Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2):110–141, 1986.

Zeev Maoz, Lesley G. Terris, Ranan D. Kuperman, and Ilan Talmud. What is the enemy of my enemy? Causes and consequences of imbalanced international relations, 1816–2001. *Journal of Politics*, 69(1):100–115, 2007.

Peter V. Marsden. Network data and measurement. *Annual Review of Sociology*, 16: 435–463, 1990.

Friedemann Mattern. Virtual time and global states of distributed systems. In *Proc. Workshop on Parallel and Distributed Algorithms*, pages 215–226, 1989.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In P. Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, 1973.

Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

Guy Melançon and Arnaud Sallaberry. Edge metrics for visual graph analytics: A comparative study. In *Proc of the 12th Intl Conf on Information Visualization*, pages 610–615, 2008.

Sigurd Meldal, Sriram Sankar, and James Vera. Exploiting locality in maintaining potential causality. In *Proc. ACM PODC*, pages 231–239, 1991.

Stanley Milgram. The small world problem. *Psychology Today*, 1(1):61–67, 1967.

Kazuo Misue, Peter Eades, Wei Lai, and Kozo Sugiyama. Layout adjustment and the mental map. *Journal of Visual Languages and Computing*, 6(2):183–210, 1995.

James Moody, Daniel A. McFarland, and Skye Bender-DeMoll. Dynamic network visualization. *American Journal of Sociology*, 110(4):1206–1241, 2005.

Jakob L. Moreno. *Who Shall Survive? Foundations of Sociometry, Group Psychotherapy, and Sociodrama*. Beacon House, 1953. Originally published in 1934.

Martina Morris. Overview of network survey designs. In Martina Morris, editor, *Network Epidemiology: A handbook for survey design and data collection*, pages 8–24. Oxford University Press, 1st edition, 2004.

Lankeshwara Munasinghe and Ryutaro Ichise. Time aware index for link prediction in social networks. In *Data Warehousing and Knowledge Discovery*, volume 6862 of *LNCS*, pages 342–353. Springer, 2011.

Keiko Nakao and A. Kimball Romney. Longitudinal approach to subgroup formation: re-analysis of newcomb's fraternity data. *Social Networks*, 15(2):109–131, 1993.

Keith V. Nesbitt and Carsten Friedrich. Applying gestalt principles to animated visualizations of network data. In *Proc. Information Visualisation*, pages 737–743, 2002.

Theodore M. Newcomb. *The Acquaintance Process*. New York: Holt, Rinehart & Winston, 1961.

Mark E.J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

Bobo Nick, Conrad Lee, Pádraig Cunningham, and Ulrik Brandes. Simmelian backbones: Amplifying hidden homophily in facebook networks. *Proc. IEEE/ACM Intl. Conf. on Advances in Social Network Analysis and Mining*, pages 525–532, 2013.

Cydney B. Nielsen, Shaun D. Jackman, Inanç Birol, and Steven J. M. Jones. Abyss-explorer: Visualizing genome sequence assemblies. *IEEE Transactions on Visualization and Computer Graphics*, 15:881–888, 2009.

Peter Nordlie. A longitudinal study of interpersonal attraction in a natural group setting. Unpublished Ph.D. dissertation, University of Michigan, 1958.

Eugene P. Odum. *Fundamentals of Ecology*. Saunders, 3rd edition, 1971.

Pietro Panzarasa, Tore Opsahl, and Kathleen Carley. Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *American Society for Information Science and Technology*, 60(5):911–932, 2009.

D. Stott Parker Jr., Gerald J. Popek, Gerard Rudisin, Allen Stoughton, Bruce J. Walker, Evelyn Walton, Johanna M. Chow, David A. Edwards, Stephen Kiser, and Charles Kline. Detection of mutual inconsistency in distributed systems. *IEEE Trans. Softw. Eng.*, 9(3):240–247, 1983.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Adam Perer and Ben Shneiderman. Balancing systematic and flexible exploration of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 12 (5):693–700, 2006.

Filippo Radicchi, Claudio Castellano, Federico Cecconi, Vittorio Loreto, and Domenico Parisi. Defining and identifying communities in networks. *PNAS*, 101 (9):2658–2663, 2004.

Ruth M. Ripley, Tom A.B. Snijders, and Paulina Preciado. Manual for RSiena, 2012. University of Oxford, Department of Statistics; Nuffield College.

Garry Robins, Pip Pattison, Yuval Kalish, and Dean Lusher. An introduction to exponential random graph (p*) models for social networks. *Social Networks*, 29(2): 173–191, 2007a.

Garry Robins, Tom A.B. Snijders, Peng Wang, Mark Handcock, and Philippa Pattison. Recent developments in exponential random graph (p^*) models for social networks. *Social Networks*, 29:192–215, 2007b.

Garry Robins, Pip Pattison, and Peng Wang. Closure, connectivity and degree distributions: Exponential random graph (p*) models for directed social networks. *Social Networks*, 31(2):105–117, 2009.

John Scott. *Social Network Analysis: A Handbook*. SAGE, 2nd edition, 2000.

Robert Sedgewick. *Algorithms in C*. Addison-Wesley, 1998.

Cosma R. Shalizi and Andrew C.Thomas. Homophily and contagion are generically confounded in observational social network studies. *Sociological Methods and Research*, 40(2):211–239, 2011.

Georg Simmel. Individual and society. In Kurt H. Wolf, editor, *The Sociology of Georg Simmel*. Free Press, 1950a.

Georg Simmel. *The Sociology of Georg Simmel*. Kurt H. Wolff (editor), Free Press, 1950b.

Mukesh Singhal and Ajay Kshemkalyani. An efficient implementation of vector clocks. *Inf. Process. Lett.*, 43(1):47–52, 1992.

Tom A.B. Snijders. Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology*, 21:149–172, 1996.

Tom A.B. Snijders. The statistical evaluation of social network dynamics. In Michael E. Sobel and Mark P. Becker, editors, *Sociological Methodology*. Basil Blackwell, 2001.

Tom A.B. Snijders. Markov chain Monte Carlo estimation in exponential random graph models. *Journal of Social Structure*, 3(2):1–40, 2002a.

Tom A.B. Snijders. The Statistical Evaluation of Social Network Dynamics. *Sociological methodology*, 31(1):361–395, 2002b.

Tom A.B. Snijders. Models for longitudinal network data. In Peter J. Carrington, John Scott, and Stanley Wasserman, editors, *Models and methods in social network analysis*, pages 215–247. Cambridge University Press, 2005.

Tom A.B. Snijders. Statistical models for social networks. *Annual Review of Sociology*, 37(1):131–153, 2011.

Tom A.B. Snijders, Philippa E. Pattison, Garry L. Robins, and Mark S. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, 36(1):99–153, 2006.

Tom A.B. Snijders, Johan Koskinen, and Michael Schweinberger. Maximum likelihood estimation for social network dynamics. *Annals of Applied Statistics*, 4:567–588, 2010a.

Tom A.B. Snijders, Gerhard G. van de Bunt, and Christian E.G. Steglich. Introduction to stochastic actor-based models for network dynamics. *Social Networks*, 32(1):44–60, 2010b.

Christoph Stadtfeld. *Events in social networks : a stochastic actor-oriented framework for dynamic event processes in social networks*. KIT Scientific Publishing, 2012.

Astrid Steffen. Decision-making in the (stressed) brain. Dissertation, University of Konstanz, KOPS, 2010. URL `http://nbn-resolving.de/urn:nbn:de:bsz:352-opus-126600`.

Christian E.G. Steglich, Tom A.B. Snijders, and Michael Pearson. Dynamic networks and behavior: Separating selection from influence. *Sociological Methodology*, 40(1): 329–393, 2010.

Klaus Stein, René Wegener, and Christoph Schlieder. Pixel-oriented visualization of change in social networks. In *Advances in Social Networks Analysis and Mining*, pages 233–240, 2010.

Robert J. Sternberg. *Cognitive Psychology*. Wadsworth Publishing, 5th edition, 2008.

Yizhou Sun, Jiawei Han, Charu C. Aggarwal, and Nitesh V. Chawla. When will it happen? — relationship prediction in heterogeneous information networks. In *Proc. 5th ACM intl. conf. on Web Search and Data Mining*, pages 663–672, 2012.

Justin Talbot, John Gerth, and Pat Hanrahan. An empirical model of slope ratio comparisons. *IEEE Transactions on Visualization and Computer Graphics*, 18(12): 2613–2620, 2012.

Sidharth Thakur and Theresa-Marie Rhyne. Data vases: 2D and 3D plots for visualizing multiple time series. In *Proc. of the 5th International Symposium on Advances in Visual Computing: Part II*, pages 929–938, 2009.

Anna Tikhonova, Carlos D. Correa, and Kwan-Liu Ma. Visualization by proxy: A novel framework for deferred interaction with volume data. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1551–1559, 2010.

Francisco J. Torres-Rojas and Mustaque Ahamad. Plausible clocks: constant size logical clocks for distributed systems. *Distrib. Comput.*, 12(4):179–195, 1999.

Melanie Tory and Torsten Moller. Evaluating visualizations: Do expert reviews work? *IEEE Computer Graphics and Applications*, 25:8–11, 2005.

Mark Trappmann, Hans J. Hummell, and Wolfgang Sodeur. *Strukturanalyse sozialer Netzwerke*. VS Verlag, 2nd edition, 2011.

Amanda L. Traud, Eric D. Kelsic, Peter J. Mucha, and Mason A. Porter. Comparing community structure to characteristics in online collegiate social networks. *SIAM Review*, 53(3):526–543, 2011.

Amanda L. Traud, Peter J. Mucha, and Mason A. Porter. Social structure of Facebook networks. *Physica A*, 391(16):4165–4180, 2012.

Edward R. Tufte. *Envisioning Information*. Graphics Press, 1990.

Edward R. Tufte. *Beautiful Evidence*. Graphics Press, 2006.

Tomasz Tylenda, Ralitsa Angelova, and Srikanta Bedathur. Towards time-aware link prediction in evolving social networks. In *Proc. 3rd Workshop on Social Network Mining and Analysis*, pages 9:1–9:10, 2009.

Matthew O. Ward. A taxonomy of glyph placement strategies for multidimensional data visualization. *Information Visualization*, 1(3/4):194–210, 2002.

Colin Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, 2nd edition, 2004.

Stanley Wasserman and Katherine Faust. *Social Network Analysis: Methods and Applications*. Cambridge University Press, 1994.

Stanley Wasserman and Philippa Pattison. Logit models and logistic regressions for social networks: I. an introduction to markov random graphs and p*. *Psychometrika*, 60:401–425, 1996.

Martin Wattenberg. Visual exploration of multivariate graphs. In *Proc. ACM SIGCHI Conference on Human Factors in Computing Systems (CHI 2006)*, pages 811–819, 2006.

Duncan J. Watts and Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, 1998.

Barry Wellman and Milena Gulia. The network basis of social support: A network is more than the sum of its ties. In Barry Wellman, editor, *Networks in a Global Village. Life in Contemporary Communities*, pages 83–117. Westview Press, 1999.

Max Wertheimer. Untersuchungen zur Lehre von der Gestalt II. *Psychologische Forschung*, 4:301–350, 1923.

Harrison C. White, Scott A. Boorman, and Ronald L. Breiger. Social structure from multiple networks. I. Blockmodels of roles and positions. *The American Journal of Sociology*, 81(4):730–780, 1976.

Ji-Soo Yi, Niklas Elmqvist, and Seungyoon Lee. TimeMatrix: Visualizing temporal social networks using interactive matrix-based visualizations. *International Journal of Human-Computer Interaction*, 26(11 & 12):1031–1051, 2010.

Tom Yokum. Sparklines: The Tom Thumb of Statistical Graphs. *Foresight: The International Journal of Applied Forecasting*, 14:48–50, 2009.