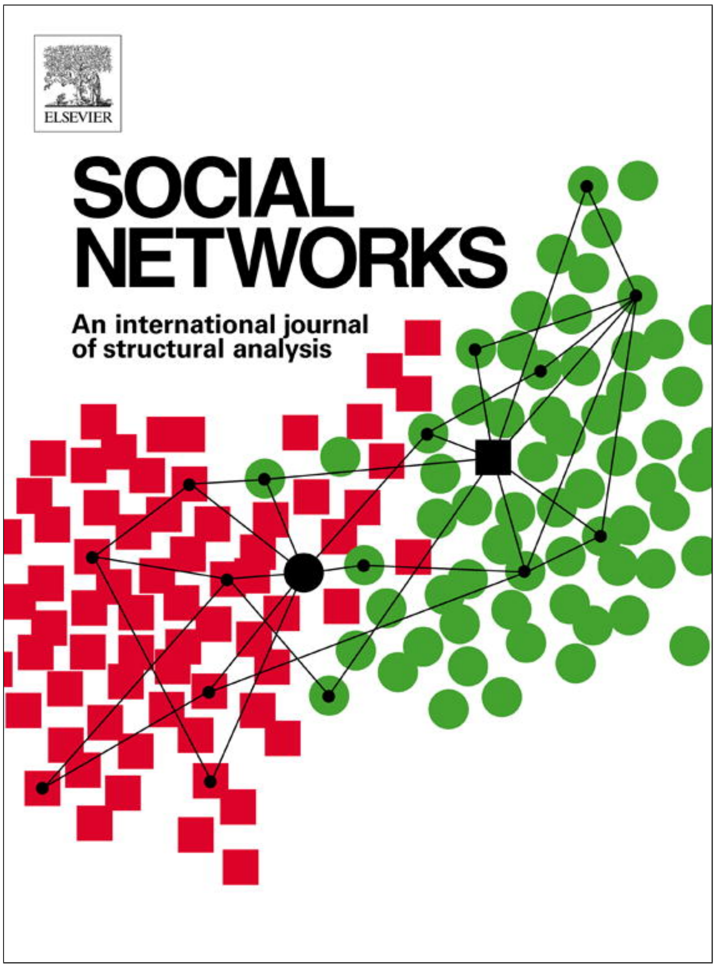
ELSEVIER
SOCIAL
NETWORKS
An international journal
of structural analysis

Social Networks 38 (2014) 50–62

Contents lists available at ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Information communities: The network structure of communication

Peter Pal Zubcsek[a,*], Imran Chowdhury[b], Zsolt Katona[c]

[a] Warrington College of Business Administration, University of Florida, Gainesville, FL 32611, United States
[b] Lubin School of Business, Pace University, New York, NY 10038, United States
[c] Haas School of Business, University of California at Berkeley, Berkeley, CA 94720-1900, United States

ARTICLE INFO

Keywords:
Communities
Communication
Information transmission
Network closure

ABSTRACT

This study puts forward a variable clique overlap model for identifying information communities, or potentially overlapping subgroups of network actors among whom reinforced independent links ensure efficient communication. We posit that the average intensity of communication between related individuals in information communities is greater than in other areas of the network. Empirical tests show that the variable clique overlap model is indeed more effective in identifying groups of individuals that have strong internal relationships in communication networks relative to prior cohesive subgroup models; the pathways generated by such an arrangement of connections are particularly robust against disruptions of information transmission. Our findings extend the scope of network closure effects proposed by other researchers working with communication networks using social network methods and approaches, a tradition which emphasizes ties between organizations, groups, individuals, and the external environment.

1. Introduction

Recent research in organizational theory has attempted to understand how individuals and groups are linked via the network structure, and how this relates to important behavioral outcomes such as collaboration and information transmission. Some scholars have attempted to explore this structure more deeply, focusing on the pattern of connections within and between groups in social networks (Everett and Borgatti, 1998; Moody and White, 2003), the nature of resources which inhere in these connections (Burt, 1992; Coleman, 1988), average path lengths between individuals and their relationship to information transmission (Killworth and Bernard, 1978), and by examining the high local clustering of individuals which seems to be a recurring characteristic of linked global communities (Uzzi and Spiro, 2005; Watts and Strogatz, 1998).

Nevertheless, our understanding of networks of individuals remains incomplete. Much remains unexplored in our structural understanding of social networks, and there is a paucity of research on how clusters of individuals actually link to each other and to the broader organizations and institutions within which they are embedded to make information transmission happen. While we know a good deal about the role of intermediaries in spreading information between disjoint groups (Burt, 1992), we know much less about the mechanisms which underlie the broader structure of information transmission within social networks (Stinchcombe, 1990). Accordingly, our paper moves to explore these mechanisms further. We do so by building on the extensive literature on the detection of cohesive subsets in social networks, which uses formal mathematical methods to define structural concepts within social networks (Borgatti et al., 1990; Forsyth and Katz, 1946; Moreno, 1934; Wasserman and Faust, 1994).

Taking up this approach, we first define a family of cohesive subgroup models called *information communities*, which consist of potentially overlapping subgroups of network actors among whom reinforced independent links facilitate efficient communication (defined, in general, as lossless information transmission). Subsequently, we demonstrate how information communities include the models of Luce and Perry (1949) and Palla et al. (2005), and introduce a new *variable clique overlap model* within this family of cohesive subgroup models. This model is distinct from prior models in that it allows researchers to more loosely (that is, they exhibit higher global connectivity) or tightly (they exhibit lower global connectivity) define cohesive subgroups as required by the question and the independent and dependent variables under study.

The form of organization we study is one that utilizes multiple independent paths throughout the network to mutually reinforce

* Corresponding author. Tel.: +1 3522733283.
E-mail addresses: pzubcsek@ufl.edu (P.P. Zubcsek), ichowdhury@pace.edu (I. Chowdhury), zskatona@haas.berkeley.edu (Z. Katona).

http://dx.doi.org/10.1016/j.socnet.2014.02.001
0378-8733/

both the strength of relationships as well as the integrity of the information traveling within a particular group. In this instance, having reinforced independent communication paths for information transmission takes on heightened importance. Imagine, for instance, the remarkable resilience of the regime of Bashar al-Assad in Syria. As of this writing, Assad and most of his government have survived an uprising that began peacefully in March 2011 and, following a harsh and repressive response by Assad's forces, quickly transformed itself into an armed uprising supported financially and militarily by numerous international opponents. While the support of important allies, such as the Russian and Iranian governments and the Lebanese political party Hezbollah, has been important in keeping Assad's regime intact, more important may be the fact that the regime's core actors have tightly coalesced as an array of domestic and international forces has aligned against it (Bellin, 2004; Robinson, 2012). This core group, which is generally thought to include Assad, family members such as Assad's brother Maher, key senior military officials, and senior business and government executives (many of whom are drawn from Assad's Allawite religious community and Christian and Sunni elites aligned with the regime), has seen relatively few defections since the start of the conflict. Working in close coordination, this group serves as the regime's "brain," driving its survival, in particular in Syria's major urban centers, and gains against rebel forces. A similar phenomenon has been noted by observers of the North Korean regime, where information access is tightly controlled and key decision-makers cluster around the supreme leader (since late 2011 Kim Jong-un and before him his father and grandfather, respectively, Kim Jong-il and Kim Il-sung) within important elite institutions such as the National Defense Commission, the Cabinet and the Korean Workers' Party (Bermudez, 2004).

While a number of communications configurations might be possible for a top leadership group, in the instances noted above a tightly-linked group with little scope for brokerage and multiple reinforced paths for information transmission ensures that outside information and influence impacts the whole group, which in this manner remains secure against attempts to dislodge it. In addition, due to its cohesive nature, the group itself can broker between the various important constituencies in the country, including the military, business interests and the civil service. Importantly, however, it is the group which serves a brokering function across the whole network, not individual actors within the group.

Tightly-controlled authoritarian governments are one area where the network structure we describe is particularly evident, but these structures can be identified in any settings where cohesive groups play a role in regulating information flow (e.g., certain departments of corporations, government security agencies, religious orders such as the Catholic Church). Indeed, while in recent years the rise and increasing complexity of information communication technologies has seemingly complicated information exchange in fundamental ways and thereby increased the likelihood of strategic information manipulation by organizational actors (Burt, 1992; Daft and Weick, 1984; Yates, 1993; Zmud, 1990), these technologies have also concomitantly strengthened the ability of organizations to build in fault tolerance mechanisms that greatly increase both the efficiency of information transmission and the likelihood that individual-level brokering behaviors will be detected and sanctioned. Thus, notwithstanding the diversity of relationships present in contemporary organizational settings, by focusing on one key characteristic of social networks, information conductivity, we may be able to increase our understanding of how the structure of connections within a particular social network is related to the survival of the organization or group within which this network is constituted.

In the next section, we briefly describe prior work in the organization studies literature which looks at communication, in particular work which examines the structure of cohesive subgroups within social networks. Section 3 introduces the generic concept of information communities and provides the mathematical details of a variable clique overlap model for identifying communities building on the research described in Section 2. In Section 4, we test the efficacy of the variable clique overlap model against earlier cohesive subgroup models on two sets of communication data – a telephone network and an organizational email network – and present our empirical methods and results. We conclude with a summary of our findings, suggestions for future research, and potential applications in Section 5.

2. Theoretical background

Starting at least as far back as Forsyth and Katz (1946), who developed the concept of a "sociomatrix," organizational scholars have noted the impact of subsets of the network which are characterized by greater cohesiveness relative to the rest of the network. Cohesive subgroups provide a crucial link between individuals and organizations – between the micro and macro levels of analysis – and are characterized by a high number of ties between individuals within the group. They are also relatively closed to outsiders, as most of the interactions of the subgroup as a whole happen between members (Borgatti et al., 1990; Freeman, 1992; Moody and White, 2003).

Early research on cohesive subgroups attempted to elucidate the mechanisms by which group behavior within social networks affected different outcomes. Subsequent research in this area has explored network structure using graph-theoretic criteria to examine group behavior (Alba, 1973; Luce, 1950; Mokken, 1979; Seidman and Foster, 1978). A working assumption of this school of thought is that an optimally cohesive subgroup is a clique, in which all subgroup members interact with one another. Cliques are important to understanding the concept of network closure. As noted by Burt (2005), networks in which people are very highly connected to each other, that is, where two actors are both connected to the same third-parties, are better at transmitting information. As the strength of third-party ties connecting two people increases, the network around them becomes more closed (Burt, 2005). Thus, closure in an organizational setting is measured by the strength of the indirect connections between individuals with colleagues acting as third parties. In this schema, some individuals are more strongly connected through third parties than others in the study population. The relationships of such individuals are said to be strongly embedded in the closed network. One of the important outcomes of strongly embedded close relationships is an increase in trust between individuals, which can lead to increased information transfer as well (Coleman, 1990).

While structural definitions of groups based on cliques have proven fairly effective at identifying various organizational variables such as relationship intensity, group centrality, and even performance (Balkundi and Harrison, 2006; Borgatti et al., 2009; Evans, 2010), further refinements are necessary to link these set definitions with particular behavioral outcomes such as interpersonal communication. A first step towards accomplishing this was introduced by Borgatti et al. (1990), who proposed using independent paths as a way to identify cohesive subsets. Specifically, they defined subgroups based on high connectivity between any pair of within-group actors. While this method results in groups – termed "Lambda sets" – that are likely to persist despite the loss of a few relationships within them, it produces non-overlapping groups which remain relatively independent from the rest of the network. One recent approach, put forward by Moody and White (2003), addresses this issue by proxying vertex connectivity (the minimum number of actors that one has to remove from a group to disconnect it) for the structural cohesion of groups. This method of analyzing vertex-independent paths between actors (no two of which can be simultaneously disconnected by the removal of one intermediary) is closely related to the variable clique overlap method for locating information communities detailed in the next section. However, the method we outline is less dependent on intermediaries in information transmission. It instead relies on the interactions of cliques themselves to predict communication intensity.

Focusing on independent communication pathways is important for several reasons. Because these pathways go through different organizational actors, removing one or perhaps even a few actors will not result in the breakdown of the pathway. Rather, the presence of alternate communication routes ensures that group cohesion is maintained. Also, because communications can flow through multiple paths, it is difficult for any one or a few actors to limit information sharing in any substantial way (Moody and White, 2003). At a structural level, this information acquisition phenomenon is also related to the idea of the clique. Individuals' informal social relations tie them into relatively cohesive subgroupings, which possess their own norms and values, and which may run counter to the formal social structure of the organization (or other social grouping) within which they are found. Cliques are often among the most important sources of a person's identity and sense of belonging and have the potential to strengthen relations between individuals (Scott, 2006). The presence of a third (or fourth, or fifth, etc.) party can curb disagreement and provide a basis for reaching consensus as a means for maintaining harmony within the group (Krackhardt, 1999).

In the next section, we describe prior models of cohesive groups and develop a variable clique overlap model for communication contexts. Before, doing this, however, we offer the following caveat: looking specifically at the context of communication networks, it is important to bear in mind that connection structure provides only a tiny fraction of the important information contained in social interactions between network actors. The nature of information exchanged is also very important, and relationships are affected by the kinds of information transmitted between two actors (e.g., whether this information is positive or negative). Thus, similar structural patterns may perhaps lead to different organizational outcomes if the content of the relationship is taken into account. Keeping in mind these considerations regarding the content of ties between individuals and groups, we nevertheless proceed in subsequent sections of this paper with a relatively structural analysis. While we do not discount the importance of the nature and content of the ties connecting actors in a given social network, the limitations of our data preclude such an analysis at present.

3. Model

In this section we review the relationship between information flow and the network structure as discussed by network researchers. We then introduce the concept of information communities and provide a variable clique overlap model for identifying communities building on prior research. Table 1 offers a summary of the cohesive group models (defined over unweighted networks) discussed in this section, including our own variable clique overlap model.

3.1. Communication and the conductivity of relationships

Our view of personal influence focuses on the information transmitting ability (*conductivity*) of relationships. We assume that if from two actors, v_1 and v_2, v_1 possesses some information, then for some $0 \leq \delta_{v_1 v_2} < 1$, v_2 obtains the same information with probability $\delta_{v_1 v_2}$. We do not focus on the dynamics of the flow of information, and we assume that all communication happens within a single examined period of time.

Under this probabilistic view and with these assumptions, higher conductivity can be understood as increased *fault tolerance* to errors in information transfer. Furthermore, since we do not generally possess information revealing the nature of ties (in terms of our model, we do not know the magnitude of particular δ-s), fault tolerance can only be associated with redundancy. Therefore, we require information communities to be structures with several independent paths between any pair of associated actors. Thus, our probabilistic assumptions relate very closely to the fundamentals of brokerage and closure: when efficient communication is a key measure of success in the organization, the most stable cohesive groups are those with high connectivity. Conversely, our model does not allow actors to be in a brokering position within information communities. Along these lines, to completely exclude the possibility of brokerage, we only consider mutual relationships to be in communities.

Network closure effects peak in the Luce and Perry cliques (1949) – maximal subgroups of actors in which all individuals know each other and among whom all choices are mutual (Wasserman and Faust, 1994). In the sociology literature, the term *clique* refers to maximal structures – in mathematical terms, maximal cliques. It is thus natural to treat the structure of cliques in a network as a community structure (see Rowley et al., 2004 for an empirical study). The main limitation of this approach, however, is that forcing so much within-group homogeneity results in very small groups of actors. For instance, whereas there are social networks with large cliques, the maximum clique size in typical communication networks does not exceed 15–30 actors.

3.1.1. Cohesive groups based on group diameter

The vast literature on cohesive subgroups offers various methods to address this problem. Herein we restrict our attention to those approaches that are applicable in a setting where the researcher only observes the structure but not the intensity of communication relationships and has to conclude the structure of cohesive groups based only on the unweighted network. The earliest such approaches focused on groups with large within-group actor similarity. Building on the notion of groups defined by Alba (1973), Mokken (1979) introduced a family of structures determined by minimum within-group distance, defining an n-clique of a network as a maximal subset of actors such that no two actors within the group are further than n steps away from each other in the network. In this schema, an n-clan is an n-clique with the restriction that any two actors sharing group membership can be connected via a path of length up to n that is contained entirely within the group. Finally, an n-club is simply a maximal subnetwork with diameter at most n. As in social networks, attributes of related actors tend to be positively correlated, restricting within-group actor distance results in homogeneous groups.

3.1.2. Cohesive groups based on within-group degree

Seidman and Foster (1978) and Seidman (1983) took a different approach. They defined groups based on minimum within-group degree. In a k-plex, the minimum within-group degree is the size of the group minus k. In a k-core, the minimum within-group degree is k. Both of these methods detect groups of high connectivity and within-group similarity. However, it is interesting to note that, unlike cliques and the other groups we discuss above, the k-cores cannot partially overlap. Instead, they form a hierarchical clustering of the actors into similarity groups.

3.1.3. Minimum across-group communication

The methods referred to thus far tend to produce structures that exhibit high within-group and low across-group connectivity.

Table 1
Overview of cohesive group models defined over unweighted networks.

Community model	Organizing principle	Overlap	References
Cliques	Baseline model	Yes	Luce and Perry (1949)
N-cliques	Maximum within-group	Yes	Mokken (1979)
N-clans	distance	Yes	Alba (1973), Mokken (1979)
N-clubs		Yes	Mokken (1979)
K-plexes	Minimum within-group	Yes	Seidman and Foster (1978)
K-cores	degree	No	Seidman (1983)
Girvan–Newman partitions	Minimum across-group	No	Girvan and Newman (2002)
Modularity-based communities	communication	No	Newman (2006)
Lambda Sets	Independent paths	No	Borgatti et al. (1990)
K-components	within groups	Yes	Moody and White (2003)
"UCINET" clique clustering	Reinforced	No	Everett and Borgatti (1998)
Uniform communities	independent paths	Yes	Palla et al. (2005)
Variable overlap communities		Yes	This study

Girvan and Newman (2002) propose that these properties should serve as a primary means of classifying actors in a network. They provide a hierarchical clustering of actors by starting from the full network and removing edges in reverse order according to their betweenness, recomputing the betweenness values after every step. However, their model has high computational requirements and is therefore impractical for analyzing large networks. Newman (2006) presents a related but faster approach: his "method of optimal modularity" also starts from the full network, but in every step it splits the network into two groups that maximize "the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random". An interesting property of this algorithm is that for every network, it generates a unique distribution of non-overlapping groups; however, as a result, smaller communities are not always identified (Fortunato and Barthélemy, 2007).

3.1.4. *Independent paths within groups*

In an influential paper building on prior work from the engineering discipline, Borgatti et al. (1990) introduced the idea of defining subgroups based on high connectivity between any pair of within-group actors. For their Lambda sets, any two actors in the same group are connected by a higher number of edge-independent paths than there are edge independent paths between any member of the group and any non-member. The so-arising groups tend not only to have both small diameter and high connectivity, but much more importantly, they are difficult to disconnect by removing only a few relationships. Thus, Lambda sets have high fault tolerance against the breaking down of a few relationships within the group. This is a very similar property to what we require for information communities. However, the method of Borgatti et al. also produces a hierarchical clustering of nodes in the network, resulting in non-overlapping groups. Further, as Moody and White (2003) point out, around actors with high brokering power, high edge connectivity still may be accompanied by low vertex connectivity. To address this issue, Moody and White define structural cohesion in networks by vertex connectivity: the (overlapping) cohesive groups are the k-components of the network. A k-component in a network is an induced subnetwork that remains connected even after removing $k-1$ actors from the group. This definition provides stronger redundancy within groups, which makes Moody and White's (2003) model particularly relevant for application in communication networks.

3.1.5. *Reinforced independent communication channels*

Whereas vertex connectivity ensures that information has multiple redundant ways to travel between actors in a group, these paths may be arbitrarily long. When the relationships in the network are imperfect at transmitting information, such long paths may cause the breakdown of the system even without the removal of actors from the group due to the degradation of information integrity in the message. To address this issue, Everett and Borgatti (1998) proposed a method of grouping actors by first identifying the Luce and Perry cliques in the network and then clustering these cliques based on the size of their overlap (number of actors shared in common). From the perspective of information transmission, these structures not only have the high vertex connectivity desired, but as the independent paths connecting within-group actors run through overlapping cliques, the information may be reinforced at every step. A specific algorithm of this kind, generating non-overlapping groups of actors, is implemented in a follow-up work by Borgatti et al. (2002). Similarly, Palla et al. (2005), drawing from work in the physical sciences, also define communities as connected collections of cliques. Their novel idea is to apply strictly local conditions – specifically, one on the size of the overlap – to decide whether two cliques belong to the same subgroup (community) or not. As a result, any two cliques sharing the required overlap are grouped together even in the case when they are linked more strongly to two, otherwise disconnected groups. This approach produces overlapping subgroups of actors. However, it does not explicitly look at (the intensity of) interactions between actors within the network, and as a result, the analysis is still highly structural and static.

3.2. *From cliques to communities*

While all models of cohesive groups described offer researchers an opportunity to examine communication within social networks, not all of them are optimal for detecting the most important communication channels within these networks. For instance, some of these approaches enforce too much within-group homogeneity (e.g., k-cores), while others allow individuals to be in brokering positions within the derived structures (e.g., modularity-based communities). Models such as the k-components and others within the family of those utilizing reinforced independent communication channels (described in Section 3.1.5) are the exception in this regard, but, as pointed out in the foregoing discussion, the k-components' reliance on long information–transmission paths may severely degrade the integrity of transmitted information, thereby decreasing the ability of the group as a whole to broker across the entire network. At the same time, structures built from cliques allow actors within these cliques (including those at the center and the periphery of a particular grouping) to access information available to the rest of the network without resorting to brokers and without their being excluded from the information flow. In this sense, such structures democratize information sharing within their particular sub-groupings.

This is an important consideration in settings, such as the examples of the Syrian and North Korean regimes mentioned earlier, where the (regulated) flow of information within a tightly connected grouping of individuals helps to ensure its survival and potentially protects it from internal and external threats by allowing rapid identification and sanctioning of behavior which violates group norms. A recent example will help illustrate this point. In December 2013, the regime of Kim Jong-un in North Korea dismissed Jang Song-thaek, Kim's uncle and the unofficial second-in-command for the country. Jang was subsequently executed, along with several of his followers in the power structure, and his other supporters in the regime are in the process of being identified and removed. In an unusual move, the Kim regime announced two detailed press releases within the period of a week outlining the reasons for Jang's removal.[1] The primary motivation for Jang's removal from power and execution seems to have been his attempts to develop his own, independent power base by going outside of the official, hierarchical organizational structures and communication channels established by Kim.

To reflect the above argument, we extend our probabilistic model of communication presented in Section 3.1 by assuming that information travels over sequences of cliques between its source and its destination. This simplification is natural since in many cases when the source of information is a single actor, the most closely related actors immediately obtain the same information. Thus, we assume that at the beginning of the examined time period, all members of some clique have some common information, to be spread in the form of messages. Under this assumption, identifying structures that are capable of efficiently transmitting this message reduces to the task of identifying those cliques (if any) which are likely to obtain the same message.

3.3. Information communities – a general model

Based on the ideas of Everett and Borgatti (1998), Palla et al. (2005) proposed that communities be understood primarily as collections of cliques as opposed to being collections of actors (whereas the mapping from cliques to actors is trivial). They defined two cliques to be adjacent when they share at least c nodes, c being a parameter that they empirically calibrated. Finally, they took the connected components of the so-built clique-network to form their communities. Below we generalize their model. We use a similar method to identify information communities but we determine the adjacency of two cliques by comparing the output of a more general clique overlap function against an adjacency threshold. The clique overlap function in our model may take into account not only the size of the overlap but also the size of the two overlapping cliques, thereby providing a more flexible necessary condition on clique overlap. As a result, we arrive to a general model that, on the one hand, incorporates the Luce and Perry cliques and the communities of Palla et al. (2005), while on the other hand offers an opportunity to better characterize situations where cohesive groups play a role in regulating information flow.

As smaller structures (due to the smaller sample size) tend to carry greater variance with respect to their properties of interest,[2] we introduce two ways that allow researchers to further refine the necessary conditions for a set of actors to be qualified as a cohesive subgroup under our method. First, smaller cliques may be excluded from the clique-network. Second, we allow researchers to specify the minimum number of actors required in a community. Thus, our general model of information communities has four parameters: the clique overlap function, the adjacency threshold, the minimum clique size and the minimum community size.

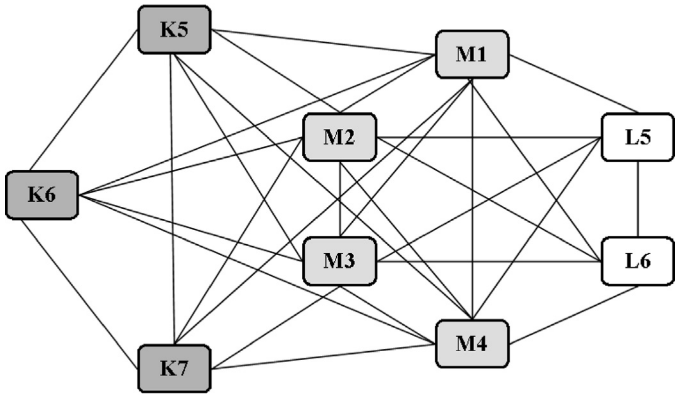

Fig. 1. Illustrating clique overlap for $k = 7$, $l = 6$, $m = 4$: the central 4 nodes are shared by the k- and l-cliques.

Definition. Let $p, q \in \mathbb{N}$ be the parameters for minimum community size and minimum clique size, respectively. Let $r \in \mathbb{R}$ be the adjacency threshold and $f(k, l, m) : \mathbb{N} \times \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ the clique overlap function. Let further S denote a subset of the nodes in a network. We say that S is an *information community* if it is a maximal structure for which

- $|S| \geq p$,
- every node in S is contained in a clique of size of at least q,
- for every pair of cliques in S, there is a series of cliques connecting them so that for consecutive cliques of size k and l having an overlap of size m, we have $f(k, l, m) > r$.

Note that by setting $f(k, l, m) \equiv 0$, $p = q = 3$ and $r = 1$, we obtain a community structure which is equivalent to the set of Luce and Perry cliques in the network. Naturally, more interesting structures can also be generated by this model. The communities of Palla et al. arise by taking $f(k, l, m) \equiv m$ (and $r = c - 2$ to exclude cliques of size less than c). As the minimum required overlap between two cliques does not depend on the size of the cliques, we refer to this model as that of "uniform" communities. This method is clearly able to discover larger social groups. However, the enforced homogeneity results in rigid structures. Below, we discuss how to better capture the underlying communication based solely on the structure of communication relationships and derive another clique overlap function $f(k, l, m)$ that, under our probabilistic model of communication, can identify communities of higher information conductivity.

3.4. A community model of variable clique overlap

In Section 3.2, we proposed that structures capable of efficient information transmission are those that can successfully share information between adjacent member cliques. To get a better understanding of the role of the function f in this process, we now analyze the role of clique overlap in the transmission of information within a network. Imagine that two cliques, K and L of sizes k and l, respectively, have m common nodes (see Fig. 1). Assume that the members of the k-clique know some information that they can transmit to the members of the l-clique, with some small (average) probability δ per link. For a node $v \in L \setminus K$, the probability of v

[1] "Report on Enlarged Meeting of Political Bureau of Central Committee of WPK" (December 9, 2013); "Traitor Jang Song Thaek Executed" (December 13. 2013), Korean Central News Agency (http://www.kcna.co.jp/index-e.htm).

[2] In the empirical section, we focus on communication frequency as the dependent measure.

receiving the message is[3]:

$$Pr[\text{message reaches } v] = 1 - Pr[\text{message does not reach } v]$$
$$\approx 1 - (1-\delta)^m \approx 1 - (1 - m\delta) = m\delta.$$

Thus, we may conclude that $f(k, l, m)$ has to be linear in m. However, whereas a small clique overlap can provide a strong medium for information exchange between two small cliques, it may be insufficient between two large cliques. To account for this, we have to normalize the value m. We therefore transform our measure to decrease in the sizes of the non-overlapping parts of the cliques, which are $k-m$ and $l-m$, respectively. The normalizing effect of having more nodes standing out in one of the cliques should be independent from the number of nodes standing out on the other side. Therefore it is natural to take $|(K\bigcup L)\setminus(K\bigcap L)| = (k+l-2m)$ to be the normalizing factor. In this work, to achieve asymptotic properties that are easier to interpret, we take a monotone transformation of $m/(k+l-2m)$ that brings the above suggested measure down to a fixed scale – we use $m/(k+l)$.

For a further refinement, we may observe that if $k \gg l$ or $l \gg k$, the transmission of the message is only efficient in one of the directions, whereas we want to identify structures of high information conductivity independent from the location of the message source. Thus, to penalize for asymmetry, we introduce the term $2kl/(k^2+l^2)$ which decreases in the absolute difference of k and l. Finally, for convenience we normalize the measure to be in [0, 1) by doubling the numerator.[4]

In sum, we set

$$f(k, l, m) \equiv \frac{4klm}{(k+l)\cdot(k^2+l^2)}.$$

Thus, the size of required overlap between two cliques to be considered adjacent varies depending on the size of the two cliques in question. We therefore refer to this model of communities as that of "variable (clique) overlap". In the next section, we empirically test the efficacy of this model by statistically comparing it to all other models of cohesive groups discussed above.

4. Empirical studies

Can cohesive groups characterize information flows in communication networks? To answer this question, we constructed a test to compare the model of variable clique overlap to other models of cohesive subgroups discussed in Section 3. We did this by exploring two communication networks: (1) a dataset of approximately 70,000 subscribers of a fixed-line telephone provider in Eastern Europe; and (2) a two-year collection of email data from the now-defunct Enron Corporation of Houston, Texas, USA. The first (telephone network) dataset provides a general communication context for exploring these phenomena; the second (email) dataset provides a specifically organizational context which may be more likely to display some of the patterns and characteristics we describe for the informational communities family of cohesive subgroup models and our variable clique overlap model within this family.

Below we summarize the common methodological details. Data-specific information and results are provided in the sections corresponding to the particular studies. We conclude by discussing the potential interpretations of our findings.

4.1. Methods and variables

In Section 3.2 we described the general framework of information communities and detailed three specific subgroup models within the so-defined family of models. These are the set of all Luce and Perry cliques, the uniform community model of Palla et al. (2005), and our approach, which defines communities by variable clique overlap. Including the further methods detailed in Section 3, Table 1 summarizes the cohesive group models that we tested in our empirical studies.

In the analysis that follows, we took the approach of Burt (2005) and used the structure of relationships in the network to predict the strength of the same relationships, proxied by the intensity of communication between related actors. The idea behind taking this route is that if it is indeed not individual actors within the group but the group itself which serves a brokering function across the whole network, then one should observe heightened communication intensity between related actors within the groups defined by the subgroup model in question. Therefore, we compared our model of variable clique overlap with prior methods of subgroup analysis in two ways. First, we took the ratio of the corresponding relationship intensities of within-structure edges and all other links, and compared these ratios across community models. Second, for every relationship in the network, we took centrality measures of the related actors plus variables describing the extent to which the two actors share membership in cohesive groups defined by the model under study. Using these measures as independent variables, we then employed a regression model to predict the intensity of the given relationship as described in Section 4.1.2.

For any model we considered, we let the term *community distribution* correspond to the set of communities identified by a certain parameterization of the model. For every such community distribution, we defined *intra-edges* as relationships in the network that are between actors who share at least one community membership. Finally, we considered every relationship that is not an intra-edge to be an *inter-edge*. In other words, two related actors who do not share any community membership are related through an inter-edge.

In both studies, we started from datasets containing communication records. Using these data, we built our communication network, defining the set of actors and the structure of relationships as detailed in the corresponding subsections below. We then computed the set of communities corresponding to various parameterizations of the general models detailed in Section 3. Subsequently, we computed the average intra-edge intensity to the average inter-edge Intensity Ratio for all of the community distributions to report the range of these values. Finally, we constructed two group membership variables and, controlling for common measures of network centrality, we statistically gauged the relationship between group membership and communication intensity.

4.1.1. Intensity Ratio

Our first method simply assigns a positive number to any community distribution the following way: *Intensity Ratio* is the ratio of the average relationship intensity of intra-edges to the average relationship intensity of inter-edges. That is, if A is the set of actors and $X \subset A \times A$ the relationships in the network, and $i(x_{a_1a_2})$ is the intensity of the relationship between the actors $a_1, a_2 \in A$, our measure *Intensity Ratio* becomes

$$\text{Intensity Ratio} = \frac{\sum_{x\in W} i(x)/|W|}{\sum_{x\in B} i(x)/|B|},$$

[3] The probability that the message spreads to some members of the target clique indirectly through interconnecting actors does not change this analysis significantly for the values of δ that we consider, and therefore we omit the corresponding terms.

[4] We have $0 \le m/(k+l) < 1/2$ by the fact that neither clique is contained in the other and $0 < 2kl/(k^2+l^2) \le 1$ by the inequality between the geometric and arithmetic means.

where W denotes the set of intra-edges and $B = X \setminus W$ that of the inter-edges.[5]

The quantity Intensity Ratio measures the extent to which the community distribution identifies the key communication links in the network. In short, the greater number of relationships characterized by intense communication within groups, the higher the value of Intensity Ratio. However, as the measure is not adjusted for the number of nodes in groups, it is often biased towards very small communities that capture only the most central region of the network. This problem is nevertheless remedied in our second method of analysis.

4.1.2. Relationship-level predictions

For a more comprehensive analysis, we employed regression models to characterize the association between membership in cohesive groups and the intensity of particular relationships. For every relationship in the network, we computed variables that reflect joint membership of the two related actors in one or more cohesive subgroups defined by the method of subgroup analysis in question. Further, we computed some standard network centrality measures that are related to the level of network closure (clustering) and the possibility for individual brokerage (degree, betweenness and reach centralities). Controlling for these measures of network centrality, we subsequently estimated the intensity of every relationship in the network within a statistical model. Comparing the coefficients and the overall regression fit allowed us to rank the studied models of subgroup analysis. Below we first briefly describe our independent variables, then discuss issues related to model selection.

4.1.2.1. Group membership. For any community distribution with W being the set of intra-edges, for any relationship $e(v_1, v_2)$, we let

$$\text{Group membership} = \begin{cases} 1 & \text{if } e \in W\text{, and} \\ 0 & \text{otherwise.} \end{cases}$$

4.1.2.2. Group overlap. Everett and Borgatti (1998) proposed ‘clique overlap centrality’ to describe the prominence of actors embedded in social networks. We took this idea to the level of relationships. Keeping the notation from above, for a given community distribution, z denoting the number of subgroups shared by the related actors v_1 and v_2, we let

$$\textit{Group overlap} = \begin{cases} z-1 & \text{if } e \in W\text{, and} \\ 0 & \text{otherwise.} \end{cases}$$

For models producing disjoint cohesive subgroups, Group overlap $\equiv 0$ and therefore did not enter the regressions.

4.1.2.3. Degree centrality. For the regression analysis predicting relationship intensity, it is most natural to control for the degrees of actors (Freeman, 1979). As implemented in UCINET, we take both in- and out-degrees of both the source and target actors in every directed relationship. For our undirected communication network, we only used one degree variable per actor.

4.1.2.4. Clustering. The clustering coefficient measures the density of connections in the network induced by the neighbors of the focal actor (Watts and Strogatz, 1998). For directed networks, our clustering measure was the density of the directed network induced by the out-neighbors of the focal actor (the measure is 1 if and only if all connections are present in both directions).

4.1.2.5. Betweenness centrality. Freeman (1977, 1979) defined betweenness centrality capturing the extent to which information in the network flows through actors. Individuals with high betweenness occupy critical positions that allows them to broker information, potentially leading to higher status. We used UCINET to compute the betweenness values of the source and target actor of every relationship.

4.1.2.6. Reach centrality. Out-reach centrality is a measure that reflects how many actors in the network information coming from the focal actor may reach in a few steps (Borgatti, 2003). In-reach, on the other hand, captures the extent to which the focal actor may be reached by information originated elsewhere in the network. We used UCINET to compute both reach centrality values for both the source and the target of our communication links.

4.1.3. Model selection

Many of the models examined in this article map multiple community distributions to one network, and often (except for the deterministic method of Newman, 2006) even the first publications of subgroup models compared herein are silent on how the researcher should best set the parameters of the given model. To treat the nondeterminism in certain models from Table 1, we considered multiple parameterizations where applicable and computed our variables for each of the so-derived community distributions. In a similar vein, instead of strictly defining the minimum number of actors required to be in a cohesive subgroup, we allowed that parameter to vary as well. For Intensity Ratio, we varied minimum cohesive subgroup size between 2 and 100, and report the minimum and maximum Intensity Ratio taken at any of the so-generated community distributions. Thus, this method characterizes each model of cohesive subgroups by an interval. Whereas we are primarily interested in the maximum of this interval (or the best-case Intensity Ratio of each model), the lower end of the interval may carry important information on how each model may perform under different parameterizations.

For the relationship-level regression models, we only manipulated the structural parameters where applicable (but not the minimum size of admitted cohesive subgroups). For the k-components, k-cores, k-plexes and n-clans we varied the connectivity/group diameter parameter between 2 and the maximum for which the routine still completes (within 48 h on a standard PC with 2.4 GHz processor frequency) and returns a non-empty community distribution. For the Lambda sets and the clique clustering of Everett and Borgatti (1998), we simply took the output of UCINET and tested for every distribution present in the output matrix. For the information communities models (Luce and Perry, 1949; Palla et al., 2005 and our variable clique overlap model) we varied q and always set the parameter r so that cliques of size less than q could not meet the overlap criterion. For the model of uniform communities, we achieved this by setting $r = q - 2$, while in our own model, we set $((q-1)/q) - 0.0001 \le r < (q-1)/q$.[6] We estimated the regression for all the so-generated models and report the regression with the highest model fit according to McFadden's pseudo-R^2.

[5] If none or all of the edges are within communities, this measure is not defined. However, distinguishing between the discussed community models on such networks lies beyond our interest.

[6] As the input of our computer algorithm, we could specify r up to four decimal digits. Since any clique in the examined networks contains less than 100 actors, our condition also implied $(q-2)/(q-1) < (q-1)/q - 0.0001 \le r$.

Table 2
Summary statistics for networks generated from the Call Traffic Dataset.

	Full network	Degree threshold		
		100	50	25
Actors	66,814	64,038	58,618	42,757
Relationships	906,752	498,482	295,485	93,546
Degree[a]	27.14 (64.20)	15.57 (13.00)	10.08 (6.86)	4.38 (2.67)
Clustering[a]	0.085 (0.064)	0.056 (0.060)	0.050 (0.062)	0.041 (0.073)
Relationship intensity[a,b]	562.07 (2955.78)	689.30 (3472.36)	807.16 (3828.06)	960.29 (4261.01)

[a] Mean values reported, standard deviation given in brackets.
[b] Total call duration, in seconds.

4.2. Study 1 – Call Traffic Data of a telephone network

To define our first communications network, we took call records from a small Eastern European fixed-line telephone provider. The dataset spans 3 months and contains individual call records of about 70,000 customers (at the node level, however, we do not possess information distinguishing private and business subscribers). Each data record contains the identifier of the calling and called parties, the duration of the phone call in seconds, plus some marketing variables such as price and price discounts for some calls. Since the latter information is not available for every record, we do not use it in our analysis.

The communication network was constructed the following way: telephone subscribers became actors and to each call we generated a relationship between the calling and the called parties. Importantly, to every call we can associate an information flow both from the calling to the called party and vice versa. Therefore, we undirected every relationship. To each pair of related actors we then assigned the cumulated call duration (in seconds, during the three months considered, in both directions) as the weight representing the intensity of the relationship between them.[7] Next, we dropped the isolated actors corresponding to subscribers who remained inactive during the period of analysis. This left 66,814 actors in the network, spanning 906,752 undirected relationships. However, the degree distribution revealed that the lack of information identifying business subscribers may indeed be a problem: there were several dozens of nodes with degrees over 1000. As such high numbers of different partners contacted within three months is unlikely for individual subscribers, we attempted to correct this error by removing nodes with the highest degrees. As this approach is clearly limited in its ability, we performed the cut for threshold degrees 100, 50 and 25. Herein we report the results for the network obtained by removing nodes of degree more than 50 but we note that all our findings are essentially the same for both the other two networks generated this way as well as for the full network of 66,814 actors.

Table 2 compares the full network to those achieved by removing the high-degree actors. After removing the actors with degree more than 50, there were 58,618 actors and 295,485 undirected relationships remaining in the dataset. The average relationship intensity over these links (call duration, in seconds) was 807.16.[8]

4.2.1. Results

Due to specifics of the algorithmic implementation, in this study, we restricted our focus to information communities.[9] Table 3 reports the range of the Intensity Ratio measure for the three models of information communities. The maximum clique size in the network is 6, and for communities only composed of cliques of size at least 5, the three models achieved a very similar performance on this metric. For smaller clique sizes, the Luce and Perry cliques obtained a higher Intensity Ratio, indicating that the tighter groups better reflected the structure of the core communication links.

Table 3
Intensity Ratios for various community models in the Call Traffic Dataset.

Model	Cliques	Uniform overlap	Variable overlap
Intensity Ratio	2.39–3.03	2.06–3.03	2.24–3.03

To understand how detecting cohesive groups may help organizational researchers in identifying the important communication channels in the network, we also conducted a more robust follow-up test. We developed a statistical model that uses the group membership variables defined in Section 4.1.2 to predict the communication frequency for each relationship. For the maximum validity of this test, we controlled for the degree and the clustering coefficient of the related actors. Due to the computational limitations mentioned above, we omitted the betweenness and reach centralities from this analysis.

Concerning the distribution of our dependent variable, existing theory does not provide us with strict guidance. Therefore, we estimated the relationship between network variables and cumulated call duration using multiple functional forms. Since the no-relationships were observed in the relationship structure and excluded from the regression, the distribution of our dependent variable was positive and continuous. The log-normal distribution is commonly used to model such distributions. Therefore, herein we report the detailed results of a node-specific fixed-effects OLS regression predicting the logarithm of cumulated call duration, formulated as:

$$\log Y_{ij} = \beta \mathrm{X}_{ij} + \xi_i + \varepsilon_{ij},$$

where $\mathbf{X}_{ij}$ includes a constant term, the group variables and the average degree and clustering coefficients of the related actors (see Section 4.1.2 for details). We further assumed that $\varepsilon_{ij} \sim N(0, \sigma^2)$ and estimated both σ and the node-specific fixed effects ξ_i inside the regression. Finally, we note that all of our findings reported below are reproduced by other functional forms, for instance estimating a Poisson regression to predict relationship intensity.

Table 4 shows the results of this analysis. All three models achieved the maximum predictive power at minimum clique size of 5; the results in the table concern these model instances. Not surprisingly, the model achieving the highest fit was the clique model of Luce and Perry (1949). Of the two other information community models, our model of variable overlap outperformed the uniform communities model in this test. However, the low model fit values indicate that there is great individual heterogeneity that neither models of cohesive groups, nor centrality measures were able to account for. We find it particularly interesting that in this context our overlap measure had the highest predictive power for the

[7] We chose the additive relation between call durations and relationship intensity to keep our analysis as parsimonious as possible.
[8] The total duration of the analyzed calls thus exceeded 7.56 years.
[9] We implemented our algorithms representing the networks as bitvectors but other algorithms store the networks via the full adjacency matrix. We stress here that this is a technicality and not necessarily an indicator of superior performance.

Table 4
Study 1: predicting communication intensity in the Call Traffic Dataset. [Fixed-effects OLS regression. Standardized coefficients reported.]

Variable	Cliques	Uniform overlap	Variable overlap	Centrality only
Degree (average)	−0.174*** (0.004)	−0.134*** (0.004)	−0.150*** (0.004)	−0.011** (0.004)
Clustering (average)	−0.025*** (0.004)	0.030*** (0.004)	−0.001 (0.004)	0.233*** (0.004)
Group membership	0.349*** (0.003)	0.392*** (0.003)	0.392*** (0.003)	
Group overlap	0.252*** (0.003)	−0.005 (0.002)	0.136*** (0.002)	
Constant	4.944*** (0.002)	4.944*** (0.002)	4.944*** (0.002)	4.944*** (0.002)
Observations	590970	590970	590970	590970
Individuals	58618	58618	58618	58618
ρ (fraction of variance due to individual fixed effects)	0.2244	0.2220	0.2225	0.2156
Pseudo R^2	0.0629	0.0372	0.0474	0.0131

model of cliques. This indicates that in networks with large individual heterogeneity, high overlap centrality may correspond to more diverse social resources, leading to stronger relationships.

4.3. Study 2 – Enron Email Dataset

To define our second communications network, we accessed the publicly available Enron Email Dataset (Federal Energy Regulatory Commission [FERC], 2003). This database contains 200,869 records of emails and allows the generation of reports related to specific data queries. Independent of email content, we converted the data into a directed graph. Actors in the database (as senders or recipients of emails) were defined as nodes. Subsequently, for every email an edge was generated from senders to all intended recipients. In this context, emails sent to distribution lists were bypassed. To every directed relationship we assigned the frequency of communication along that link as a weight.

Since our measures are naturally defined on directed networks, we kept the asymmetry of the communication relationships.

However, as in the directed case our communities are built from maximal *directed cliques*, only those actors who were both senders and recipients of emails could be in communities. Therefore, only nodes with the above property were extracted for analysis. Thus, only nodes exhibiting this property were extracted for analysis. This did not weaken our results since on the one hand, the communication via the affected links contained plenty of spam messages, while on the other hand – possibly due to prior data cleansing efforts of the FERC – the communication involving actors who were only senders or only recipients of emails still remained at a much lower average intensity than the average intensity in the remaining network (reported below).

We converted strings of data into names of individuals, storing trivial solutions in a relational database. Remaining strings were matched to available information, and new nodes were created when no match was possible. This method allowed us to identify approximately 6000 individuals within the data set. Subsequently, we matched the recipients of emails to those individuals in our database, reducing the set of nodes to 3455 individuals. The number of induced directed relationships contained therein was found to be 50,931, at an average intensity (frequency of communication in the direction of the relationship) of 7.51. In accordance with the objectives of our research, we did not perform dichotomizations of the relationships. Thus, in our analyses, the minimum link intensity was 1, while the maximum intensity was 676. Table 5 reports some descriptive statistics of the so-generated network.

It is important to note that while there were a total of almost 3500 nodes in the final network we constructed, a number of key organizational actors – including the one-time CFO Andrew Fastow – were not part of this dataset. This is due to a number of factors, including the fact that some senior executives bypassed emails as a major method of communication, probably to avoid leaving a written record of potentially-sensitive communications. While certainly this is one limitation to the data, this sample does include information for former Chairman and CEO Kenneth Lay, an important organizational actor whose prominence in the network, as calculated by standard centrality measures, was above the median for the data set.

Table 5
Descriptive statistics of the Enron Email network.

	Mean	St. dev.
Out-degree	110.67	(619.60)
In-degree	110.67	(279.60)
Clustering	0.19	(0.18)
Out-reach	1032.25	(165.58)
In-reach	1032.25	(142.86)
Betweenness	8765.46	(50768.00)
Relationship intensity[a]	7.51	(22.09)

[a] Number of emails.

4.3.1. Results

The computation of the K-plexes and the communities based on edge betweenness did not terminate within 48 hours of starting the computation process. We therefore omitted these models from the analysis. For the N-clans, the only computation that terminated within 48 hours was that for $N = 2$.

Table 6 reports the Intensity Ratios for various community models. The results offer two important things insights. First, we can see that the only models that reached double-digit ratios are the Lambda sets of Borgatti et al. (1990) and the clique clustering method of Everett and Borgatti (1998) (the variable overlap model we propose achieved the third highest values at this test). Second, the community models defining structural cohesion in terms of independent paths obtained higher Intensity Ratios than the rest of the models tested.

To study how much help certain models offer at *detecting* the most active communication links in the network, we again used a statistical model to predict communication intensity: the number of emails between the related actors in each relationship. To obtain robust results, we controlled for the degree, the clustering coefficient, the betweenness and reach centralities of related actors. As

Table 6
Intensity Ratios for various community models in the Enron Email Dataset.

Model	Cliques	Uniform overlap	Variable overlap
Intensity Ratio	1.97–4.43	1.90–4.64	1.94–6.29

Model	UCINET clustering	K-components	K-cores
Intensity Ratio	1.74–23.92	1.83–3.69	1.35–3.69

Model	Lambda sets	Modularity groups	N-clans[a]
Intensity Ratio	1.81–15.48	1.59–1.61	1.74–1.89

[a] Computationally feasible only for $N = 2$.

our dependent measure is the count of emails within a given time period, we modeled the data by estimating a Poisson regression.

We let Y_{ij} denote the number of emails in our sample sent by actor i to actor j. Since we only included pairs of actors between whom there was at least one email sent (when the communication relationships are observed, this information is given to the organizational researcher), we used $Y_{ij} - 1$ as dependent variable in the Poisson regression. (As $(Y_{ij} - 1 | Y_{ij} \geq 0)$ has the exact same distribution as Y_{ij}, the Poisson link function is still theoretically correct in this case.) In sum, we formulated the Poisson regression the following way:

$$\log(E(Y_{ij} - 1 | Y_{ij} \geq 1, \mathrm{X_{ij}})) = \beta \mathrm{X_{ij}},$$

where $\mathbf{X}_{ij}$ includes a constant term plus the network variables enlisted in Section 4.1.2. We estimated the model using the maximum likelihood method.

Table 7 reports the results of this regression. In this study, the network variables better explained communication patterns: including only the network centrality measures achieved a pseudo R^2 almost 0.35. The models of cohesive groups discussed in this paper improved the predictions further: the best-performing variable overlap model achieved a pseudo R^2-value above 0.41. It is also clear that methods considering overlapping cliques can be expected to perform better than those models that did not require the presence of independent paths with frequent information reinforcement along them.

Estimating the Poisson regression for various parameterizations of the models considered allows us to offer recommendations concerning the optimal values of these parameters. For our variable overlap model, the maximum fit was achieved at minimum clique size of $q = 5$ (and $r = 0.7999$). Also, for any minimum clique size below 7, our model performed better than all relaxations of the Luce and Perry cliques considered herein. Thus, considering that the variable overlap model also achieved the best fit for minimum clique size 5 over the Call Traffic Data, setting $3 \leq q \leq 6$ and $0.6666 \leq r \leq 0.8333$ seems to be a good starting point when applying our model over other communication networks. Concerning the rest of the models of cohesive groups, we highlight several interesting findings below. It is surprising that whereas the cliques also have the strongest predictive power for minimum size of 5, the uniform communities predicted communication best at the minimum clique size of 8. The k-cores performed best at $k = 10$, and the k-components at $k = 11$. The latter fact is particularly interesting since at such a high connectivity value, the communities detected by the method of Moody and White (2003) did not overlap, whereas for lower values of k the Overlap variable provided an extra predictor in the statistical system we applied.

5. Discussion and conclusion

In this paper, we developed and tested a variable clique overlap model for identifying information communities, or potentially overlapping subgroups of network actors among whom reinforced independent paths facilitate efficient communication (lossless information transmission). Empirical tests in two distinct contexts – a telephone network and an organizational email network – validated the model as a useful tool for studying information transmission within communication networks and compared well with earlier models examining the impact of group structure on organizational outcomes. In addition, we found that our model had greater predictive power in the organizational email network than in the more general telephone communication network. Our findings suggest that new insights can be gained by grouping maximal cliques of individuals within social networks based on the degree to which they overlap; the pathways for information transmission generated by such an arrangement of connections are particularly robust against disruptions.

5.1. Contributions

This paper provides useful information for both organizational scholars and practitioners. On the methodological side, our model can be utilized by network researches as a means to identify overlapping community structures in communication networks. Breaking with the traditional clustering approach, the method which we detail focuses instead on the set of cohesive groups in a social network and the interactions of those groups with each other. This is achieved by capturing the flow of information in the network at various levels – at nodes, at cliques, and at the level of communities – and allows researchers to examine contexts where the potential benefits of closure exceed those obtained from brokerage.

Our twofold contribution builds upon prior work on cohesive subgroups to define the *information community* family of models which incorporate a clique-based conception of groups within social networks (i.e., Luce and Perry cliques and the communities of Palla et al., 2005). In addition, we generalize these earlier models by incorporating a more flexible *variable clique overlap function* to provide the basis of clustering cliques in the network. Since our model is particularly relevant for studying communication networks, we examined settings where network closure was hypothesized to have an impact on information transmission. By focusing directly on the structure of primary ties between individuals, we tested the efficacy of our variable clique overlap model in the context of communication networks. We demonstrated that this conception of information communities provides researchers a powerful tool for identifying groups of individuals that have particularly strong internal relationships in closed social networks.

The test of the variable clique overlap model in communication networks also confirmed earlier work which posits a link between network closure and relationship intensity in social networks. By extending this link to the communication network setting, however, we demonstrated the extent to which different models of cohesive subgroups are able to identify relationships which may serve as important communication channels in an organizational context.

We further demonstrated a novel way to differentiate social networks from non-social networks, and how the pure mathematical machinery of network analysis used in natural sciences such as physics should be refined when analyzing social networks, in particular those found within organizations. We did this by examining contexts where cohesive groups may play an active role in regulating the flow of information in the network. Indeed, our definition of information communities is based on information transmission between cliques in the network. The set of Luce and Perry cliques and the uniform communities of Palla et al. (2005) are both special cases of this general model.

When we empirically compared the performance of our variable clique overlap model against these two existing models of information communities, we found that the variable overlap model was significantly better at identifying core links in the communication network than Palla et al.'s (2005) uniform communities model. In addition, it was often better than the set of Luce and Perry cliques in the network in this regard. Further, as the information communities identified by the variable overlap model can be of larger size than the biggest clique in the network, we showed how to extend the basic concepts of network closure to large-scale networks.

These insights can be related to a number of important emergent organizational phenomena with related practical implications. Building on the mechanisms outlined with the example of authoritarian political regimes presented earlier, we can, for instance, envision overlapping networks of cliques as "centers of

Table 7
Study 2: predicting communication intensity in the Enron Email Dataset. [Poisson regression. Standardized coefficients and robust standard errors (adjusted for correlation at each source node) reported.]

Variable	Cliques	Uniform overlap	Variable overlap	UCINET clustering	Centrality only
Source out-degree	0.392*** (0.021)	0.390*** (0.019)	0.398*** (0.021)	0.382*** (0.020)	0.357*** (0.023)
Source in-degree	−0.084* (0.033)	−0.006 (0.036)	−0.082* (0.035)	0.007 (0.035)	0.104* (0.043)
Target out-degree	−0.002 (0.020)	−0.000 (0.018)	0.012 (0.020)	0.007 (0.019)	0.015 (0.019)
Target in-degree	0.292*** (0.043)	0.330*** (0.041)	0.275*** (0.040)	0.321*** (0.037)	0.432*** (0.047)
Source clustering	−0.170** (0.064)	−0.181** (0.064)	−0.168** (0.064)	−0.160* (0.064)	−0.088 (0.066)
Target clustering	0.065* (0.031)	0.005 (0.026)	0.032 (0.028)	0.055 (0.028)	0.113*** (0.032)
Source betweenness	−0.051 (0.049)	−0.061 (0.050)	−0.054 (0.051)	−0.062 (0.048)	−0.134* (0.055)
Target betweenness	−0.088** (0.030)	−0.070* (0.028)	−0.093** (0.029)	−0.065* (0.027)	−0.170*** (0.034)
Source out-reach	0.110 (0.084)	0.063 (0.085)	0.095 (0.083)	0.113 (0.092)	0.146 (0.088)
Source in-reach	0.066 (0.063)	0.102 (0.073)	0.049 (0.064)	0.103 (0.058)	0.163** (0.062)
Target out-reach	−0.056 (0.041)	−0.045 (0.042)	−0.088* (0.041)	0.006 (0.043)	0.143** (0.044)
Target in-reach	−0.004 (0.039)	−0.060 (0.037)	−0.017 (0.037)	−0.026 (0.035)	−0.102** (0.039)
Group membership	0.392*** (0.027)	0.361*** (0.037)	0.418*** (0.030)	0.344*** (0.023)	
Group overlap	0.074*** (0.011)	0.057*** (0.009)	0.094*** (0.010)		
Constant	1.466*** (0.048)	1.474*** (0.048)	1.456*** (0.048)	1.474*** (0.047)	1.541*** (0.047)
Observations	50931	50931	50931	50931	50931
Pseudo R^2	0.4089	0.3895	0.4152	0.3961	0.3498

Variable	*K*-components	*K*-cores	Lambda sets	Modularity groups	*N*-clans[a]
Source out-degree	0.384*** (0.020)	0.381*** (0.019)	0.392*** (0.019)	0.349*** (0.023)	0.376*** (0.021)
Source in-degree	0.014 (0.033)	0.021 (0.033)	0.046 (0.036)	0.071 (0.043)	0.024 (0.036)
Target out-degree	−0.000 (0.019)	−0.000 (0.019)	0.015 (0.019)	0.009 (0.019)	0.026 (0.018)
Target in-degree	0.303*** (0.038)	0.305*** (0.039)	0.347*** (0.041)	0.390*** (0.044)	0.331*** (0.038)
Source clustering	−0.145* (0.064)	−0.129* (0.064)	−0.127* (0.064)	−0.101 (0.067)	−0.143* (0.067)
Target clustering	0.059* (0.028)	0.058* (0.028)	0.044 (0.029)	0.107** (0.031)	0.058 (0.031)
Source betweenness	−0.057 (0.044)	−0.065 (0.046)	−0.060 (0.048)	−0.107* (0.053)	−0.081 (0.051)
Target betweenness	−0.052* (0.026)	−0.052 (0.027)	−0.045 (0.028)	−0.137*** (0.033)	−0.081** (0.027)
Source out-reach	0.073 (0.085)	0.061 (0.085)	0.069 (0.088)	0.173* (0.085)	0.115 (0.090)
Source in-reach	0.084 (0.058)	0.119 (0.068)	0.064 (0.070)	0.160* (0.063)	−0.012 (0.058)
Target out-reach	0.022 (0.041)	0.021 (0.042)	−0.043 (0.047)	0.152*** (0.043)	−0.060 (0.047)
Target in-reach	−0.079* (0.035)	−0.078* (0.036)	−0.107** (0.037)	−0.071 (0.040)	−0.110** (0.037)
Group membership	0.306*** (0.026)	0.282*** (0.030)	0.373*** (0.038)	0.170*** (0.042)	0.228*** (0.040)
Group overlap					0.230*** (0.026)
Constant	1.493*** (0.047)	1.499*** (0.047)	1.475*** (0.048)	1.521*** (0.048)	1.511*** (0.048)
Observations	50931	50931	50931	50931	50931
Pseudo R^2	0.3808	0.3757	0.3799	0.3543	0.3771

[a] For the *N*-clans, $N = 2$ for computational reasons.
* $p<.05$.
** $p<.01$.
*** $p<.001$.

legitimacy" in what are relatively unregulated internal knowledge markets within organizations. In other words, as the availability of knowledge, especially electronic knowledge, has increased in organizational life, the incidence of strategic information behaviors (Zmud, 1990) has also increased, resulting in a greater need for knowledge filters within organizations and groups (Hansen and Haas, 2001; Simon, 1997). As attention is a scarce resource in organizations, the need to allocate it wisely is of tantamount importance for organizational efficiency and effectiveness. In this respect, having a "group" of contacts which provides relatively reliable, useful information is an indispensable asset. We can imagine, for instance, the context of military groups in combat situations, which are highly dependent on receiving accurate, timely information to fulfill their missions. In such a situation, having access to a legitimate, reliable set of interlocking command structures and groupings which systematically filter extraneous information and transmit the "right" information may be a matter of life and death.

Further, we believe that being able to identify information communities using the variable clique overlap model would allow researchers to identify regions of the social network where information exchange is more likely, not only in terms of information exchanged in the case of emails or telephone calls or other contacts, but also in terms of attention given to the information by knowledge receivers. As demonstrated by Hansen and Haas (2001), in regions of the network where many documents are exchanged, more selected and concentrated providers of electronic documents received greater attention. Similarly, documents and information which have a pre-existing "signal of quality" as determined by the linkage to members of a group of overlapping cliques, may be more likely to be processed by receivers. While it is true that these documents would also probably be less likely to need processing by receivers due to the legitimacy conferred by their linkage to structurally important nodes, it is nevertheless important to consider that such associations serve an important filtering function in that they direct attention towards those documents which are associated with the most important or prominent members of the network.

Beyond the issue of information processing, organizational groups that have multiple, independent, paths of communication are more likely to share traits related to organizational culture and are thus more likely to survive even if some edges are removed (Borgatti et al., 1990). For instance, the persistence of certain organizational divisions and departments despite the fact that they may not necessarily be cost-effective for the organization may be explained by the phenomenon we highlight in this paper. These departments' links to the broader organization and the larger community may indeed legitimize their existence and persistence and make difficult their removal. Thus, organizational structure issues can be linked to issues of culture and organizational politics to explain outcomes which seem to deviate from rational economic accounts.

5.2. Limitations and implications for future research

As with other models which attempt to link structural aspects of groups within social networks to organizational outcomes, the variable clique overlap model faces the challenge of balancing the conceptual quality of the model with its real-world impact (Carley, 2002). However, this limitation also opens up several potential directions for extending this research further. First, we characterized organizational environments in this paper by examining the relative benefits of network closure over brokerage. This raises two important issues. On the one hand, discussing reliable measures and means to identify these relative benefits is very important for the direct applicability of our findings. On the other hand, if the communities we define are desirable structures within a given organizational communication system, ensuring the presence of the higher benefits from network closure within the organization imposes a very important policy making question.

Second, we concentrated on the parsimony of our work. Following the logic of Burt (2005), we identified relevant network structures based solely on the connection patterns. The generality of our approach affects the applicability of our research in several ways. A promising area of future research could examine the moderating role of communication content to our results. For instance, do implied relationships tend to be stronger when the connection patterns reflect not only the existence of communication, but also communication about a particular, narrow, topic? Further, in the organizational setting, if the model could capture demographic information, aspects of organizational hierarchy and structure, etc. it might provide better insight into not only the expected communication patterns within a given network, but also how the flow of information may be utilized at different nodes and within different clusters of individuals. Future work might attempt to identify how our methods can be refined by including such variables into the analysis.

Third our work could be applied to better identify bottlenecks within organizational communication systems. These bottlenecks would not necessarily be identified by traditional brokerage arguments: in our model brokerage does not exclusively correspond to individuals but is broadened to include groups of individuals that may play a role in regulating information flows (structurally this corresponds to the idea of overlapping cliques as we have illustrated). Beyond this, and in general, our techniques could be used to optimize the throughput of organizational communication networks. Combining ideas from the above two paragraphs, it would be natural to relate our work to knowledge flows in organizations (Cool et al., 1997; Hansen, 2002; Nahapiet and Ghoshal, 1998).

Another promising area of applications is marketing. As our methods take a step toward better estimating social influence among members of a communication network based only on the connection patterns, they can serve three fast developing areas in network marketing. First, the communities defined by our model can be interpreted as a segmentation of a networked market. It would be interesting to see future research discover how much the so-defined social groups share consumption patterns, general interests or specific knowledge. Second, our findings may help marketers identify opinion leaders in networks, facilitating viral marketing practices. Third, as social influence can lead to extra revenue flows to the firm through word of-mouth, our work should provide ground for improving the existing customer relationship management techniques.[10]

[10] These techniques evaluate customer profitability by comparing discounted revenue flows from the customer to the cost of acquisition and the cumulative cost of retention. For more details, see Bolton (1998).

Furthermore, despite using longitudinal data in our empirical studies, we constructed our networks by collapsing all communication records into one network layer. In a recent paper, Palla et al. (2007) consider shorter periods of aggregation as they focus on how communities evolve over time. This idea could be used for analyzing changes in organizational communication systems, ultimately improving their efficiency.

5.3. Conclusion

In summary, this study contributes to the important and growing literature on "structural patterning" (Kilduff and Brass, 2010) in social networks, in particular studies of clique structure and its relationship to outcomes of organizational interest. Although earlier work has suggested that overlapping cliques might offer a way to study different organizational outcomes, our study refined this concept for the communication network context. Studying variable clique overlap models allows us to get closer to fully describing communication networks with no indispensable central nodes, which have built-in redundancy mechanisms that allow the organization to function despite the potential presence of internal fissures and external disruptions. The examples noted in the preceding paragraphs are but a few where our extension of network closure theory can help the organizational researcher estimate relationship intensity in the absence of data on information traffic, solely based on connection patterns (which information is clearly easier to obtain than more detailed records on communication). Subsequent studies can test the efficacy of this model when the content of information is also taken into account.

Acknowledgments

We would like to thank Tom D'Aunno, Andras Benczur, Gokhan Ertug, Martin Gargiulo, Jennifer Petriglieri, Miklos Sarvary, Jarrett Spiro, and Barry Wellman for their helpful comments on early drafts of this manuscript. We would also like to acknowledge Social Networks Co-Editor Patrick Doreian and the two anonymous reviewers for their developmental comments during the review process.

References

Alba, R.D., 1973. A graph-theoretic definition of a sociometric clique. Journal of Mathematical Sociology 3, 113–126.

Balkundi, P., Harrison, D.A., 2006. Ties, leaders, and time in teams: Strong inference about the effects of network structure on team viability and performance. Academy of Management Journal 49 (1), 49–68.

Bellin, E., 2004. The robustness of authoritarianism in the middle east: exceptionalism in comparative perspective. Comparative Politics 36 (2), 143–144.

Bermudez, J., 2004. Information and the DPRK's Military and Power-holding Elite. HASSIG, Kongdan Oh. North Korean Policy Elites. 1ª ed. Institute of Defense Analyses, Alexandria, Junho, pp. 31.

Bolton, R.N., 1998. A dynamic model of the duration of the customer's relationship with a continuous service provider: the role of satisfaction. Marketing Science 17 (1), 46–65.

Borgatti, S.P., 2003. In: Breiger, R., Carley, K., Pattison, P. (Eds.), Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers. National Academy of Sciences Press, Ch. The Key Player Problem, pp. 241–252.

Borgatti, S.P., Everett, M.G., Freeman, L.C., 2002. UCINET for Windows: Software for Social Network Analysis. Analytic Technologies, Harvard, MA.

Borgatti, S.P., Everett, M.G., Shirey, P.R., 1990. LS sets, lambda sets and other cohesive subgroups. Social Networks 12, 337–357.

Borgatti, S.P., Mehra, A., Brass, D., Labianca, G., 2009. Network analysis in the social sciences. Science 5916, 892–895.

Burt, R., 1992. Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge, MA.

Burt, R., 2005. Brokerage and Closure. Oxford University Press, New York.

Carley, K.M., 2002. Simulating society: the tension between transparency and veridicality. In: Agent 2002 Conference, Chicago, IL, p. 2.

Coleman, J.S., 1988. Social capital in the creation of human capital. American Journal of Sociology 94, 95–120.

Coleman, J.S., 1990. Foundations of Social Theory. Bellknap Press, New York.

Cool, K.O., Dierickx, I., Szulanski, G., 1997. Diffusion of innovations within organizations: electronic switching in the bell system, 1971–1982. Organization Science 8, 543–559.
Daft, R.L., Weick, K.E., 1984. Toward a model of organizations as interpretation systems. Academy of Management Review 9 (2), 284–295.
Evans, T.S., 2010. Clique graphs and overlapping communities. Journal of Statistical Mechanics, P12037.
Everett, M.G., Borgatti, S.P., 1998. Analyzing clique overlap. Connections 21, 49–61.
FERC, 2003. Information Released in Enron Investigation. http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp
Forsyth, E., Katz, L., 1946. A matrix approach to the analysis of sociometric data: preliminary report. Sociometry 9 (4), 340–347.
Fortunato, S., Barthélemy, M., 2007. Resolution limit in community detection. Proceedings of the National Academy of Sciences 104 (1), 36–41.
Freeman, L.C., 1977. A set of measures of centrality based on betweenness. Sociometry 40, 35–41.
Freeman, L.C., 1979. Centrality in social networks: conceptual clarification. Social Networks 1, 215–239.
Freeman, L.C., 1992. The sociological concept of group: an empirical test of two models. American Journal of Sociology 98, 152–166.
Girvan, M., Newman, M.E.J., 2002. Community structure in social and biological networks. Proceedings of the National Academy of Sciences United States of America 99, 8271–8276.
Hansen, M.T., 2002. Knowledge networks: explaining effective knowledge sharing in multiunit companies. Organization Science 13, 232–248.
Hansen, M.T., Haas, M.R., 2001. Competing for attention in knowledge markets: Electronic document dissemination in a management consulting company. Administrative Science Quarterly 46, 1–28.
Kilduff, M., Brass, D.J., 2010. Organization social network research: core ideas and key debates. Academy of Management Annals 4 (1), 317–357.
Killworth, P.D., Bernard, H.R., 1978. The reverse small world experiment. Social Networks 1, 159–192.
Krackhardt, D., 1999. The ties that torture: Simmelian tie analysis in organizations. Research in the Sociology of Organizations 16, 183–210.
Luce, R.D., 1950. Connectivity and generalized cliques in sociometric group structure. Psychometrika 15, 169–190.
Luce, R.D., Perry, A., 1949. A method of matrix analysis of group structure. Psychometrika 14, 94–116.
Mokken, R.J., 1979. Cliques, clubs and clans. Quality and Quantity 13, 163–173.
Moody, J., White, D.R., 2003. Structural cohesion and embeddedness: a hierarchical concept of social groups. American Sociological Review 68 (1), 103–127.
Moreno, J.L., 1934. Psychological abstracts. In: Who Shall Survive? A New Approach to the Problem of Human Interrelations, vol. 8. Nervous and Mental Disease Publishing Co., Washington, DC, Ch.
Nahapiet, J., Ghoshal, S., 1998. Social capital, intellectual capital, and the organizational advantage. Academy of Management Review 23, 242–266.
Newman, M.E.J., 2006. Modularity and community structure in networks. Proceedings of the National Academy of Sciences 103 (23), 8577–8582.
Palla, G., Barabási, A.-L., Vicsek, T., 2007. Quantifying social group evolution. Nature 446, 664–667.
Palla, G., Derényi, I., Farkas, I., Vicsek, T., 2005. Uncovering the overlapping community structure of complex networks in nature and society. Nature 435, 814–818.
Robinson, G.E., 2012. Syria's long civil war. Current History December, 331–336.
Rowley, T.J., Baum, J.A.C., Shipilov, A.V., Greve, H.R., Rao, H., 2004. Competing in groups. Managerial and Decision Economics 25 (6–7), 453–471.
Scott, J., 2006. Social Network Analysis. Sage Publications, London.
Seidman, S., 1983. Network structure and minimum degree. Social Networks 5, 269–287.
Seidman, S., Foster, B.L., 1978. A graph-theoretic generalization of the clique concept. Journal of Mathematical Sociology 6, 139–154.
Simon, H.A., 1997. Administrative Behavior: A Study of Decision-making Processes in Administrative Organizations. Free Press, New York, NY.
Stinchcombe, A., 1990. Information and Organizations. University of California Press, Berkeley, CA.
Uzzi, B., Spiro, J., 2005. Collaboration and creativity: The small world problem. American Journal of Sociology 111, 447–504.
Wasserman, S., Faust, K., 1994. Social Network Analysis: Methods and Applications. Cambridge University Press, New York.
Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of small-world networks. Nature 393.
Yates, J., 1993. Control Through Communication: The Rise of System in American Management, vol. 6. The Johns Hopkins University Press, Baltimore, MD.
Zmud, R.W., 1990. Organizations and communication technology. In: Fulk, J., Steinfield, C. (Eds.), Opportunities for Strategic Information Manipulation Through New Information Technology. Sage Publications, Newbury Park, CA, Ch., pp. 95–116.