A Framework For Community Identification in Dynamic Social Networks

Chayant Tantipathananandh*
Dept. of Computer Science
University of Illinois at Chicago
Chicago, IL 60304
ctantipa@cs.uic.edu

Tanya Berger-Wolf*
Dept. of Computer Science
University of Illinois at Chicago
Chicago, IL 60304
tanyabw@uic.edu

David Kempe†
Dept. of Computer Science
University of Southern California
Los Angeles, CA 90089
dkempe@usc.edu

ABSTRACT

We propose frameworks and algorithms for identifying communities in social networks that change over time. Communities are intuitively characterized as "unusually densely knit" subsets of a social network. This notion becomes more problematic if the social interactions change over time. Aggregating social networks over time can radically misrepresent the existing and changing community structure. Instead, we propose an optimization-based approach for modeling dynamic community structure. We prove that finding the most explanatory community structure is NP-hard and APX-hard, and propose algorithms based on dynamic programming, exhaustive search, maximum matching, and greedy heuristics. We demonstrate empirically that the heuristics trace developments of community structure accurately for several synthetic and real-world examples.

Categories and Subject Descriptors: I.6.5 [Simulation and Modeling]: Model Development

General Terms: Algorithms.

Keywords: Dynamic Social Networks, Community Identification.

1. INTRODUCTION

Social networks are the graphs of interactions between individuals, and play an important role in the dissemination of information, innovations, or diseases. Edges can represent social interactions, organizational structures, physical proximity, or even more abstract interactions such as hyperlinks or similarity. Social networks have attracted a large amount of attention from epidemiologists [21, 25, 28], sociologists [5, 29, 38], biologists (animal interactions) [7, 6, 10, 30, 36], the intelligence community (terrorism networks) [3, 23, 24], and more recently also from computer scientists [1, 11, 14, 17, 18, 20, 22]. One of the most important questions in social networks is the identification of "communities", which are loosely defined as collections of individuals who interact unusually frequently [12, 13, 16, 18, 27, 38]. The identification of communities often reveals interesting properties shared by the members, such as common hobbies, social functions, occupations, etc. In a more general setting, including hyperlinked documents such as the WWW, these properties include related topics or common viewpoints, which has led to a large amount of research on identifying communities in the web graph or similar settings [15, 22].

In analyzing social networks, one property has until recently been largely ignored: the fact that they tend to change dynamically. When faced with dynamic social networks, most studies would either analyze a snapshot of a single point in time, or an aggregation of all interactions over a possibly large time window. Both approaches may miss important tendencies of these dynamic networks; indeed, the ongoing change of a network and its possible causes may be among the most interesting properties to observe. Consider the following simplistic scenario: individuals 1 and 2 are observed to interact at every point in time, whereas individual 3 interacts with both of them about half of the time. If the observation sequence is $\langle\{1,2\},\{1,2,3\},\{1,2\},\{1,2,3\}\rangle$, we may decide whether or not we consider individual 3 a full member of the community. On the other hand, if the sequence is $\langle\{1,2\},\{1,2\},\{1,2,3\},\{1,2,3\}\rangle$, a much more plausible explanation is that individual 3 joined the group during the observation period.

The necessity to delve into the dynamic aspects of network behavior may be clear, yet it would not be feasible without the data to support such explicitly dynamic analysis. Rapidly growing electronic networks, such as emails, web, blogs, and friendship sites, as well as mobile sensor networks on cars and animals, provide an abundance of dynamic social network data that for the first time allow the temporal component to be explicitly addressed in network analysis.

Recently, Berger-Wolf and Saia [4] proposed a framework for identifying communities in dynamic social networks, making explicit use of temporal changes. Most communities tend to evolve gradually over time (see, e.g., [2]), as opposed to assembling or disbanding spontaneously. Thus, whenever information about events in the social network is available, it is desirable to use this temporal information not only to identify communities with high intra-community similarity, but also to observe their persistence and development.

*Work supported in part by the Microsoft award 14936
†Work supported in part by NSF CAREER Award 0545855

KDD'07, August 12–15, 2007, San Jose, California, USA.

In this paper, we propose a new framework for community identification. We assume that time is discrete, and in each time step, we observe social interactions in the form of several complete subgraphs of individuals (not every individual needs to be observed in each time step), or affiliations [34, 35]. Based on these observed groupings, we want to identify "true" underlying communities and their developments over time, so that most of the observed interactions can be explained by the inferred community structure. We pose this question as a combinatorial optimization problem, based on the observation that individuals tend to (1) not change their "home community" too frequently [2], and (2) tend to interact with the home community most of the time [38, p. 320](a mathematically precise definition is given in Section 2).

After formulating the problem as an optimization problem in this way, we show that it is NP-complete and even APX-hard and present algorithms to (approximately) optimize the community structure discovered. The algorithms are based on the observation that we can separate out two subproblems: identifying which groups at time t will "become" which other groups at time $t+1$, and subsequently assigning individuals to groups. The latter part can be solved optimally using Dynamic Programming. The former part has a smaller search space, and is thus amenable to exhaustive search in small data sets. For larger data sets, we propose a set of greedy heuristics. These approaches are discussed in more detail in Sections 3 and 4.

In order to evaluate our proposed algorithms and heuristics, we consider synthetic data sets with known embedded communities, as well as two social network data sets: the well-known Southern Women data set [9, 13], capturing interactions of a small group of women in 1933 in Natchez, TN at multiple social gatherings, and the Grevy's zebra data set [36], capturing physical proximity within a zebra herd over a period of time. Our evaluation consists of two parts: (1) the framework, and (2) the algorithms. To evaluate the former, we use (inefficient) exponential-time algorithms to find the optimum solutions, which in turn necessitates a restriction to small networks. Once we establish that our framework leads to the identification of meaningful community structure, we evaluate more efficient approximate heuristics on larger real-world networks. The very encouraging experimental results are described in Section 5.

2. PRELIMINARIES

We model social networks as (undirected) graphs $G = (V, E)$. Following the motivation from social networks, the vertices V will also be called individuals, and the edges E interactions. To model dynamic interactions, we follow and slightly extend the approach of Berger-Wolf and Saia [4] and that of affiliation networks [38, Section 8]. There is a set $X = \{i_1, \ldots, i_n\}$ of *individuals*, and a sequence $H = \langle P_1, P_2, \ldots, P_T \rangle$ of *observations*. Each P_t is a collection of non-empty and pairwise disjoint sets $g_{j,t} \subseteq X$, called *groups* of individuals at time step t. The interpretation is that, for a time step t, the individuals of a group $g_{j,t}$ were observed interacting with each other, but not with the individuals of any other group $g_{j',t}$ for $j' \neq j$. Such a group may correspond to a physical or virtual gathering of its members. Notice that we do not require the P_t's to be partitions; some individuals may not be observed at all at certain times.

We stress that, in the terminology we are using, groups and communities are not necessarily the same: groups capture only a snapshot of interaction at one point in time, while communities are latent concepts which should explain many of the actual observed interactions, though not necessarily all of them.

Our framework is somewhat restrictive in the types of observations it can accommodate: only transitive interactions are allowed. While there will certainly be scenarios in which non-transitive interactions between group members happen concurrently, such scenarios would make it much more difficult to give meaningful graph-theoretic interpretations of "community", and our definition will still capture many natural real-world scenarios. Extending the allowed observations to arbitrary graphs poses an interesting direction for future research.

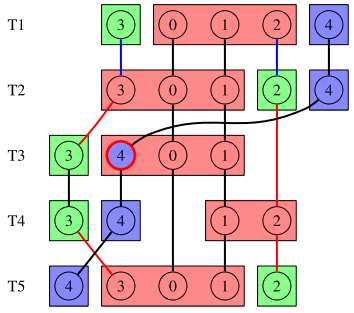


Figure 1: An example data set of 5 individuals and 6 time steps. Circles are individuals and are labeled with their IDs, rectangles are groups. Communities are identified by matching colors, and the affiliation of an individual is shown by its color.

Figure 1 shows a sample data set of 5 individuals in 6 time steps. In this data set, individuals 0 and 1 are always together in a group whenever both are present, while the other individuals take turns joining the group of the first two individuals.

2.1 Problem Formulation

Standard definitions of communities in social networks rely on various measures of cohesiveness [26; 38, Section 7.6]. Notice, however, that identifying dense subsets statically is trivial in our setting; after all, we assumed that each observation was a set of disjoint cliques. The interesting aspect of our formulation, and the one that requires novel approaches, is the temporal change in group membership.

In deriving an optimization formulation of community identification, we make the following explicit assumptions about the behavior of individuals:

1. *In each time step, every group is a representative of a distinct community. If two groups are present at the same time, there is a reason they are separate and, thus, represent distinct communities.*

2. *An individual is a member of exactly one community at any one time. While the individual can change community affiliation over time, it is affiliated with only one community at any given moment.* Notice that this does not preclude an individual from belonging to multiple communities over the course of the observation. It requires that the individual, in each time step, determines "which hat to wear today".

3. *An individual tends not to change its community affiliation very frequently.*

4. *If an individual does change its community affiliation several times, it will usually be an oscillation among a small number of communities, rather than promiscuity among many.* In other words, if an individual keeps changing its affiliations among many different communities, then it is not a true member of any of those communities.

5. *An individual is frequently present in the group representing the community with which it is affiliated. It rarely misses being with its community's group, and rarely is with other community's groups.* That is, individuals within a community interact more than those in different communities.

We will use these properties to define an optimization problem, in which we assign costs to deviation from the behaviors posited above. These costs can be intuitively modeled as a graph coloring problem (albeit with a different objective function from traditional graph coloring).

Our graph G has one *individual vertex* $v_{i,t}$ for every individual $i \in X$ and each time t. In addition, there is one *group vertex* $v_{g,t}$ for every group $g \in P_t$[1]. For each individual i and time $t \leq T-1$, there is an edge from $v_{i,t}$ to $v_{i,t+1}$. Finally, we have an edge between $v_{i,t}$ and $v_{g,t}$ whenever $i \in g$ at time t. Figure 2 shows the graph model of the example described in Figure 1. Here, the circles are individual vertices and the squares are group vertices.

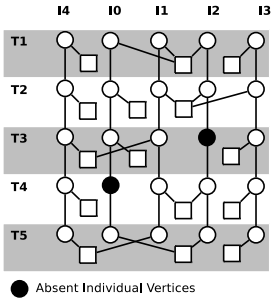


Figure 2: A sample graph model. Circles are individual vertices, and squares are group vertices. Black circles are individuals unobserved at that time.

We define a *community interpretation* of a graph G as a vertex coloring $f : V \to \mathbb{N}$ of G. The color of an individual vertex $v_{i,t}$ at time step t represents the individual i's community affiliation at time t. Similarly, the color of a group vertex $v_{g,t}$ gives the community that g represents at time t. Notice that this definition automatically ensures that each individual belongs to exactly one community in each time step, and each group represents exactly one community. We call a community interpretation *valid* if and only if, for each time step t, no two groups g and g' share the same color. This ensures Postulate 1, namely that only one group represents each community in each time step.

[1]The group vertices can also be viewed as hyperedges and the entire structure as a hypergraph.

To measure the quality of a community interpretation, we use costs to penalize violations of Postulates 3–5. There are three different types of costs, *individual*, *group*, and *color*, abbreviated as *i-cost*, *g-cost*, and *c-cost*. To allow for different relative importance of these properties, we parametrize the problem by non-negative parameters α, β_1, β_2, and γ.

i-cost: An individual cost of α is incurred whenever an individual changes its color. That is, if $f(v_{i,t}) \neq f(v_{i,t+1})$ (i.e., the individual edge does not have matching colors), a cost of α is added to the total cost.

g-cost: The group cost can be incurred for two reasons: (1) if an individual vertex does not have an edge to the group of the same color ($f(v_{i,t}) = f(v_{g,t})$, but $i \notin g$), a cost of β_1 is incurred. (2) If an individual vertex has an edge to a group of a color different from its own ($i \in g$ at time t, but $f(v_{i,t}) \neq f(v_{g,t})$), a cost of β_2 is incurred. Thus, if an individual is present at time t, but not in its group, then it incurs both costs. The first cost penalizes the individual for being absent from its current community, while the second cost penalizes the individual for being different from its current group.

c-cost: Finally, we assign a cost of γ for each color an individual uses beyond its first, so the color cost for i is $\gamma \cdot (|\{f(v_{i,t}) : t = 1, \ldots, T\}| - 1)$.[2]

The optimization problem is then to find the valid community interpretation minimizing the total cost resulting from the individual edges, group edges, and color usage. Once such a coloring f has been found, we identify each *community* C_i with the set of groups $g = g_{j,t}$ of color $f(v_{g,t}) = i$. The *community structure* is then the collection $\mathcal{C}$ of all communities. Notice that we explicitly allow a community to change or evolve over time. Once we have a community structure, we can derive from it an *affiliation sequence* for every individual i, the sequence $A_i = \langle f(v_{i,1}), \ldots, f(v_{i,T}) \rangle$ of communities that i was a member of during the observation period.

Notice that by altering the parameters $\alpha, \beta_1, \beta_2, \gamma$, we can alter the dynamic expressiveness of the model. If α and γ are large, then individuals will virtually never change group membership, and we will recover the static community structure as a special case. As α gets smaller, more frequent changes are possible, and oscillating behavior between groups can be inferred as an explanation of the observations. Finally, once γ is small as well, we will find solutions in which individuals change frequently among many groups, essentially allowing the model to accommodate frequent and complete changes of community structure. Having all parameters of the same order (in our case, all equal) turns out to provide a good tradeoff to infer "meaningful" real-world dynamic behavior in most cases.

2.2 Complexity of the Problem

Unfortunately, solving the community interpretation problem optimally is NP-complete, and cannot even be approximated arbitrarily well. We formally define the decision problem Community Interpretation as follows: Given cost

[2]Note that we could of course omit the constant -1 term, but assigning no penalty for the first color makes a manual interpretation of results more intuitive.

parameters α, β_1, β_2, and γ, a set X of n individuals and a sequence $H = \langle P_1, P_2, \ldots, P_T \rangle$ of observations, as well as an upper bound B on the total cost, is there a community interpretation for (X, H) of total cost at most B?

THEOREM 1. *The* COMMUNITY INTERPRETATION *problem is NP-complete and APX-hard. That is, there is a constant ϵ such that unless P=NP, no polynomial time algorithm can achieve an approximation guarantee better than $(1+\epsilon)$ for* COMMUNITY INTERPRETATION.

PROOF. For membership in NP, simply observe that given a certificate in the form of a coloring (community interpretation), the total cost can be computed in polynomial time by summing up all i-costs, g-costs, and c-costs, and then compared to B.

To prove NP-hardness and APX-hardness, we give an approximation preserving reduction from the MINIMUM MULTIWAY CUT problem, which is known to be APX-hard [8]. We reduce from the following APX-hard special case $k = 3$: Given an undirected graph $G = (V, E)$ with unit edge costs, and three distinct *terminal vertices* $s_1, s_2, s_3 \in V$ as well as a bound c on the total number of edges cut, is there a set C of at most c edges such that all of s_1, s_2, s_3 are disconnected from each other in the graph $(V, E \setminus C)$?

Let $n = |V|, m = |E|$. The COMMUNITY INTERPRETATION instance has an individual for each vertex v of G. During each of the first $m+1$ time steps, each of the singleton groups $\{s_1\}, \{s_2\}, \{s_3\}$ is observed (none of the other individuals are observed during those times). Let $e_1 = \{u_1, v_1\}, \ldots, e_m = \{u_m, v_m\}$ be an arbitrary ordering of the edges. During time step $m+1+t$, exactly one group $\{u_t, v_t\}$ is observed. That is, for each edge of G, the two endpoints are observed together exactly once. To complete the reduction, we set $B = c$, and $\alpha = \gamma = m + 1$, $\beta_1 = 0$, and $\beta_2 = 1$.

To prove that this is an approximation preserving reduction, we give a cost-preserving mapping from multiway cuts to valid colorings and vice versa. First, given a cut C, let S_1, S_2, S_3 be the connected components containing s_1, s_2, s_3, respectively. Give all vertices in S_i the same color in the coloring instance. For all vertices not in any of the three components (if any), arbitrarily color them with the same color as s_1. Finally, color each group of size 1 with the color of its unique member, and each group of size 2 with the color of one of its members. First, notice that this is a valid coloring, as in the first $m + 1$ time steps, all singleton groups have distinct colors, and in the remaining m time steps, there is only one group in each time step. Because individuals never change color, no i-costs or c-costs are incurred. Thus, the only cost is β_2 whenever an individual is present, but does not have the color of its group. This never happens for groups of one individual, and happens for groups of two individuals if and only if the corresponding edge e_t is cut. Thus, the total cost is exactly the same as the number of edges cut.

Conversely, if we have a valid coloring of the observation sequence, we notice that without loss of generality, it does not incur any i-costs or c-costs, since a cheaper solution (of cost m) could always be obtained by simply assigning a fixed and distinct color to each individual. Second, we notice that each of s_1, s_2, s_3 must have distinct colors. Otherwise, they would incur a total g-cost of at least $m + 1$ during the first $m+1$ steps, and again, a cheaper solution could be obtained trivially by assigning all vertices a fixed distinct color. Define S_1, S_2, S_3 to be the sets of vertices with the same color as s_1, s_2, s_3, respectively. Let S_4 be the set of all remaining vertices (if any). Notice that we can assume w.l.o.g. that all remaining vertices have the same color; otherwise, a cheaper valid coloring could be obtained by coloring all of S_4 with the same color. Let C be the set of edges cut by the partition (S_1, S_2, S_3, S_4). C by definition is a multiway cut separating s_1, s_2, s_3. Furthermore, the groups incurring a cost of $\beta_2 = 1$ in the coloring are exactly those corresponding to edges in C, as they are the ones with two distinct vertex colors, meaning that one of the vertices must have a color different from the group. (If any group of two vertices did not have the color of either of its individuals, we could obtain a cheaper solution by recoloring that group.) Thus, we have proved that the size of C is exactly equal to the cost of the group coloring we started with, completing the approximation preserving reduction.

The reduction trivially implies that the problem is NP-hard: in the decision version, we make the cost bound B equal to the bound c on the number of edges cut. □

3. FINDING OPTIMAL COLORINGS

We showed that the problem of inferring community structure in dynamic networks is NP-hard. Thus, for larger instances, we will use heuristics or approximation algorithms. However, before moving to fast heuristics, it is important to evaluate the conceptual power of the proposed approach. In other words, does the objective function defined above truly capture meaningful community structure, at least if the optimum solution is found? Certainly, if the objective function itself already misrepresented the desired intuitive notion, finding fast heuristics would be a worthless endeavor. Hence, we first present an optimal algorithm based on exhaustive search and dynamic programming. Once we have evaluated the optimal algorithm, and shown that its results are meaningful, a secondary question will be whether the solutions discovered are less informative in the real world when heuristics are used instead of an optimal algorithm.

The key observation to making an exact algorithm somewhat tractable is that once a coloring for the group vertices has been fixed, an optimum coloring for the individual vertices can be computed using Dynamic Programming. While the running time of the dynamic program is not polynomial, its only exponential dependence is 2^C, where C is the number of communities. This tends to be significantly more tractable than searching over arbitrary colorings of individuals. Since there tend to be significantly fewer group vertices than individual vertices, an exhaustive search over all group vertex colorings can be feasible for smaller instances.

At the heart of the dynamic programming approach is the observation that, given a fixed group coloring, all incurred costs can be associated with one individual at a time. That is, the cost incurred by an individual does not depend on the colors chosen for other individuals, once the group colors are fixed.

LEMMA 2. *Given a coloring of the group vertices, the minimum cost coloring of the individual vertices consists of minimum cost colorings of the vertices of each individual i, independent of other individuals.*

PROOF. The total cost is the sum of all i-costs, g-costs, and c-costs, over all relevant edges. i-costs occur only for edges $(v_{i,t}, v_{i,t+1})$, and thus only depend on the colors of

the corresponding individual i. Similarly, g-costs arise from present or absent edges $(v_{i,t}, v_{g,t})$, and thus only depend on the color of individual i, for a fixed group coloring. Finally, the c-cost for an individual i clearly depends only on the colors for that individual. The total cost is the sum over all individuals and all of their edges of the corresponding i-costs, g-costs, and c-costs, and is thus minimized if the sum is minimized for all individuals i independently. □

We will use this observation to derive an optimal algorithm for coloring individual vertices, given a group coloring. For now, assume that we have such an algorithm $\mathcal{A}$, which, given a group coloring f_g, finds the optimum individual coloring $\mathcal{A}(f_g)$ for this particular group coloring f_g. By exhaustively trying all group colorings f_g, and invoking $\mathcal{A}(f_g)$ for each of them, we ensure that we will find the optimum solution, since the optimum solution must use *some* group coloring, f_g^*, which the exhaustive search also tries.

We will describe a dynamic programming based algorithm $\mathcal{A}$ first. Next, we will give heuristics to be used in place of the exhaustive search. These heuristics will not guarantee that the best group coloring f_g^* is indeed considered, but so long as a group coloring f_g "close enough" to f_g^* is tried, they will find a good solution.

3.1 Individual Coloring

By Lemma 2, it is enough to describe an algorithm for optimally coloring any one individual over the T time steps. By running this algorithm for each individual i, we obtain a complete optimal coloring. So we fix one individual i.

Let ϵ denote the *blank* color. Let $f_g(t)$ be the color of the group in which i participates at time step t, $f_g(t) = \epsilon$ if and only if i was unobserved. By this convention, an individual with color ϵ at time step t is said to be *unaffiliated* with any communities at that time step. Let $C = \{\epsilon\} \cup \{f_g(t) : t = 1, \ldots, T\}$ be the set of all group colors of i, including the blank color ϵ. An optimal coloring for i can only use colors in C; any other color would incur a c-cost and g-cost without any compensating benefit.

Let $\Phi(t) = \{S \subseteq C : 1 \leq |S| \leq t\}$ denote the collection of all possible subsets of colors used between time step 1 and t (notice that it is possible to use a color $c \in C$ that was not assigned to a group g during time steps $1, \ldots, t$). Let $G(t, x)$ be the g-cost of coloring i at time step t with color x, $I(t, x, y)$ the i-cost of coloring i at time steps t and $t-1$ with colors x and y, and $C(x, R)$ the c-cost of using color x when R is the set of colors used in prior steps. Notice that $G(t, x), I(t, x, y)$ and $C(x, R)$ can all be easily computed given the group coloring and the parameters α, β_1, β_2, and γ.

The recurrence for the minimum cost of coloring i in time step t with color $x \in S$, having used all the colors in the set $S \in \Phi(t)$, is

$$\Gamma(t, S, x) = G(t, x) + \min_{\substack{R \in \Phi(t-1), y \in R \\ R \cup \{x\} = S}} (\Gamma(t-1, R, y) + I(t, x, y) + C(x, R)). \quad (1)$$

$$\Gamma(1, \{x\}, x) = G(1, x) \quad (2)$$

Note that if a set $R \subseteq C$ satisfies $R \cup \{x\} = S$, then R must be either S or $S \setminus \{x\}$. Hence, the search is only over two candidate subsets. In the initial condition (2), there is no cost term $C(x, \emptyset)$, as we defined the first color to be free.

Because the dynamic program explicitly optimizes over all relevant choices for time steps t and $t-1$, we obtain the following Lemma by induction:

LEMMA 3. *Given a group coloring, the optimal cost of coloring an individual i at time step t with a color $x \in S$, having used all the colors in $S \in \Phi(t)$, is given by the Equations (1) and (2).*

Applying this lemma to the end of the time horizon T, we derive the following main theorem:

THEOREM 4. *Given a group coloring, the minimum cost of coloring the individual i is*

$$\min_{S \in \Phi(T), x \in S} \Gamma(T, S, x).$$

3.2 Time and Space Complexity

In analyzing the time and space requirements of the dynamic programming approach, we first observe that we only need to retain the optimal solutions for the immediately preceding time step. Thus, the table in step t needs to have size $\sum_{k=1}^{t} k\binom{|C|}{k} = O(|C|2^{|C|})$. Of course, we implicitly still generate a table with an entry for each triple (t, S, x). Finding the optimum value for a triple (t, S, x) involves trying two different candidate sets R, each containing at most $|C|$ colors. Thus, each of the $O(T|C|2^{|C|})$ entries is computed in time $O(C)$, for a total running time of $O(T|C|^2 2^{|C|})$. Finally, the Dynamic Program is run independently for each of the individuals i, giving a total running time of $O(nT|C|^2 2^{|C|})$, with a space requirement of $O(|C|2^{|C|})$.

3.3 Group Coloring

In order to verify that the objective function captures actual community structure, we first optimize it exactly, by using exhaustive search over all group colorings. As before, we consider a group coloring as a mapping f from groups g to colors $\mathbb{N}$. We assume that the groups are indexed so that all groups at time step t precede all groups at time step $t+1$. Then, we can simply search exhaustively over all valid assignments of color to these indexed groups. We speed up the exhaustive search by using Branch-and-Bound techniques and restricting the search space by providing a limit on the total number of colors in an optimal solution.

4. GROUP COLORING HEURISTICS

To avoid the exponential time required for exploring all valid colorings, we next investigate group coloring heuristics. Once the heuristic has found a group coloring, we still apply dynamic programming to color the individuals.

4.1 Bipartite Matching Heuristic

Intuitively, a group coloring is good if most of the individuals can retain their color from one step to the next. This avoids incurring either i-costs or g-costs for those individuals. We can use this intuition to derive a matching-based heuristic. For each pair g, g' at time steps $t, t+1$, we add an edge between $v_{g,t}$ and $v_{g',t+1}$, with weight $|g \cap g'|$. Using standard flow techniques, we then find a maximum weight bipartite matching among the group vertices for those two time steps. The matching then defines which group g' "inherits" the color of g (if any). This maximizes the number of

individuals whose color of the affiliated group stays the same from t to $t+1$. Since in particular with sparse group structures (most individuals are absent), the maximum weight matching can be far from an optimal group coloring, we can augment the heuristic by enumerating all or many maximal matchings, and choosing from among them based on the actual coloring cost. The enumeration can be done efficiently, in time $O(n)$ per matching [37].

The bipartite matching heuristic intuitively aims to minimize i-costs, but does not take g-costs into account as much. Thus, it tends to perform well with fairly stable community membership. However, it tends not to pick out oscillations, as it only looks at consecutive time steps, and does not observe trends across two or more steps. Also, while maximum flow computations tend to be fairly efficient, they may not be efficient enough for large instances, in which many such computations must be performed. We therefore consider even more efficient greedy heuristics.

4.2 Greedy Heuristics

The matching algorithm tries to maximize the amount of "similarity" preserved from one time step to the next. More generally, we can define a notion of similarity for *all* pairs of groups g, g' occurring at different time steps. For concreteness, assume that the similarity is normalized to the interval $[0, 1]$, where 0 denotes disjoint groups, and 1 identical groups. Then, a good approach is to find a coloring maximizing the pairwise similarity over all pairs with the same color. While this problem is still NP-complete, we can use it as a point of departure for a class of greedy heuristics.

First, we need a notion of similarity between sets. Many measures have been proposed; a standard one in the literature is Jaccard's index [19] $\mathrm{Jac}(g, g') = \frac{|g \cap g'|}{|g \cup g'|}$, measuring the overlap of g, g' relative to the total number of elements in the sets. If we want to give more weight to similar groups in close temporal proximity, we can scale the Jaccard index by the difference in time steps. Formally, we define $\mathrm{JacD}(g, g') = \frac{\mathrm{Jac}(g,g')}{|t-t'|}$, where $t \neq t'$ are the time steps at which g and g' occur, respectively. There are several other natural similarity measures that can be combined with our approach. Due to space constraints, we defer a more detailed discussion to the full version of this paper.

With appropriate notions of similarity in place, we can define a class of greedy algorithms, iteratively assigning two groups the same color. Our algorithms are based on the idea of neighbor-joining clustering [31], and differ merely in the order in which candidate pairs of groups are considered.

In the most basic version of the greedy algorithm, we repeatedly select the pair (g, g') with the highest similarity, and decide that g, g' should have the same color. We thus grow "uni-colored components", much in the style of Kruskal's MST algorithm. Initially, each group is its own component. We then consider the edges of non-zero similarity by decreasing similarity measures. If an edge e under consideration connects two components C, C', we merge them into one component, unless they already contain groups $g \in C, g' \in C'$ such that both g, g' occur in the same time step t. In the latter case, the edge e is discarded. After all edges have been considered in this way, the algorithm terminates, and assigns each component of groups its own distinct color.

We can obtain a different greedy algorithm by proceeding in increasing order of time steps. Initially, each group at time step 1 obtains its own unique color. In iteration t, we consider all groups at time step t. We repeatedly find the edge (g', g) connecting a group g at time step t with a group g' at time step $t' < t$ of largest similarity, and color g with the same color as g', provided that no other group at time t has this color already. If g has no edge of non-zero similarity to earlier time steps, or all such colors are already taken, then g obtains its own color. Once all groups at time step t are assigned colors, we move to the next iteration. We call this algorithm the *Backward Greedy* algorithm, because it only searches backward in its color assignments

Finally, we can obtain a somewhat more restrictive version of Backward Greedy, by requiring that the edge e point no further into the past than "necessary". Specifically, when considering group g at time t, we find the latest time $t' < t$ such that g has a non-zero edge to some group g' at time t'. Then, the highest-similarity edge is selected among all edges between g and groups at time t', and g is colored accordingly, as in Backward Greedy. We call this modified algorithm *Least Delay Greedy*.

Obviously, since the problem is NP-complete, it is easy to derive examples where none of these heuristics will find the optimum solution. Furthermore, they are incomparable, in that for any pair, one can provide examples where one heuristic performs better than the other, and vice versa.

5. EXPERIMENTAL RESULTS

The main goal of our paper is to present a formal framework of community identification in dynamic networks and to show that it captures the concept of community. Thus, we first show that the optimization problem of community identification produces valid communities. We then show that the proposed heuristics result in communities similar to the optimum and thus perform well in practice. Since the problem is NP-hard, we can only find the optimal solution in small data sets. Thus, we first validate the dynamic community framework on small synthetic data sets, and also use those small data sets to compare the heuristics to the optimal algorithm. Once both the definitions and the heuristics are validated, we proceed to apply them to larger practical data sets.

We begin by inferring communities in two synthetic data sets with known embedded communities. Next, we study two real-world data sets in which communities are identified by human domain experts.

5.1 Synthetic Data sets

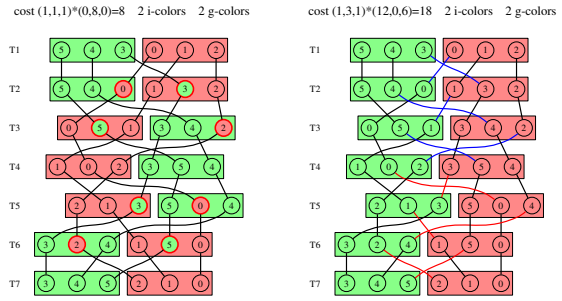
We consider two data sets in which all individuals are always present. As a result, we can safely set $\beta_1 = 0$. We consider two cost settings $(\alpha, \beta_2, \gamma) = (1, 1, 1)$ (high i-cost) and $(\alpha, \beta_2, \gamma) = (1, 3, 1)$ (high g-cost). Intuitively, these settings tend to use different explanations for switches by an individual: the former tends to posit "temporary aberrations", while the latter assumes frequent actual affiliation changes.

Assembly Line

The *Assembly Line* example models communities in which small changes happen in each time step, aggregating to complete membership changes over longer periods of time. Real-world examples include companies, PhD students in a department, casts of TV shows, or cells in the human body [33]

(as well as contents of an assembly line). The philosophical question of what "identity" of an organism means in light of replacement of all individual parts was already studied in ancient Greece, and is known as Theseus' Ship paradox [32].

An assembly line has $n = km$ individuals and m groups. In time step t, the i^{th} group consists of individuals $(ki + t)$ mod $n, \ldots, (ki+t+k-1)$ mod n. That is, in each time step, the lowest-numbered member of each group moves to the next lower group (wrapping around at n). Figure 3 shows an example of an Assembly Line with $n = 6$ individuals and $m = 2$ groups.



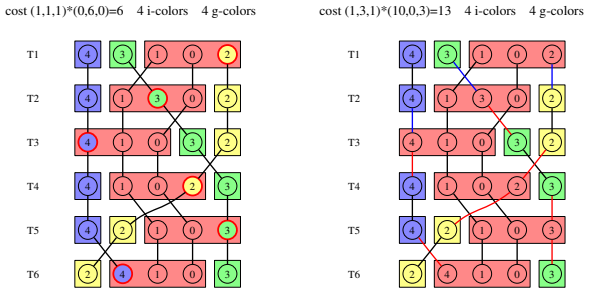
(a) (b)

Figure 3: Optimal colorings of Assembly Line with costs $(\alpha, \beta_1, \beta_2, \gamma) = (1, 0, 1, 1)$ and $(\alpha, \beta_1, \beta_2, \gamma) = (1, 0, 3, 1)$.

Figure 3(a) shows the optimal coloring under the cost setting $(\alpha, \beta_2, \gamma) = (1, 1, 1)$ in which i-cost is relatively high. Thus, individuals do not change colors, and the community a group represents is determined by a simple majority vote. In particular, the result is similar to what would be obtained by aggregating groups over time, and applying static analysis. On the other hand, with the cost setting $(\alpha, \beta_2, \gamma) = (1, 3, 1)$, the g-cost is high. Figure 3(b) shows that the resulting coloring has individuals change their community membership to match their group. Thus, the identity of groups stays the same even as the individual members change. We also note here that in this particular instance, the greedy heuristic leads to the optimal coloring for parameters $(\alpha, \beta_2, \gamma) = (1, 1, 1)$ using the Jaccard similarity measure, and for $(\alpha, \beta_2, \gamma) = (1, 3, 1)$ using the JacD measure.

Dutiful Children

Another common dynamic scenario is a population with several mostly stable communities, and a few "roaming" individuals, such as parents visiting their children in turn. In our example (Figure 4), we have three children (individuals 2,3,4), visited in turn by their parents (individuals 0,1), at times 1,4 (child 2), 2,5 (child 3), and 3,6 (child 4), respectively. The importance of this data set is that it shows a situation where the smallest number of colors needed to optimally color the graph is strictly greater than $P_{\max}$, the size of the largest partition in a time step. In Figure 4, notice that with the setting $(\alpha, \beta_2, \gamma) = (1, 1, 1)$, the solution actually recovers the "underlying" structure of a roaming pair joining three communities of individuals. For this example, we observe that the greedy algorithm with either similarity measure finds the optimum solution for both cost settings.



(a) (b)

Figure 4: Optimal colorings of Dutiful Children with costs $(\alpha, \beta_1, \beta_2, \gamma) = (1, 0, 1, 1)$ and $(\alpha, \beta_1, \beta_2, \gamma) = (1, 0, 3, 1)$.

Cost	OPT	Jaccard	
		1	$1/d$
Assembly Line			
High i-cost	**8**	**8**	18
High g-cost	**18**	20	**18**
Dutiful Children			
High i-cost	**6**	**6**	**6**
High g-cost	**13**	**13**	**13**

Table 1: Cost comparisons on synthetic data sets.

We summarize the cost of the heuristics compared to the optimal cost in Table 1.

5.2 Real-World Data Sets

Southern Women

Southern Women [9] is a data set collected in 1933 in Natchez, TN, by a group of anthropologists conducting interviews and observations over a period of 9 months. It tracks 18 women and their participation in 14 informal social events such as garden parties and card games. The event participation table is shown in Figure 5, taken verbatim from [9]. The columns, each representing an event, are not ordered chronologically, but are manually arranged by the table authors to illustrate two communities at the upper-left and lower-right corners.

Names of Participants of Group I	Code Numbers and Dates of Social Events Reported in *Old City Herald*													
	(1) 6/27	(2) 3/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 3/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 2/23	(12) 4/7	(13) 11/21	(14) 6/2
1. Mrs. Evelyn Jefferson	×	×	×	×	×	×	.	×	×	.	.	.	.	.
2. Miss Laura Mandeville	×	×	×	.	×	×	×	×	.	.	.	.	.	.
3. Miss Theresa Anderson	.	×	×	×	×	×	×	×	×	.	.	.	.	.
4. Miss Brenda Rogers	×	.	×	×	×	×	×	×	.	.	.	.	.	.
5. Miss Charlotte McDowd	.	.	×	×	×	.	×	.	.	.	.	.	.	.
6. Miss Frances Anderson	.	.	×	.	×	×	.	×	.	.	.	.	.	.
7. Miss Eleanor Nye	.	.	.	.	×	×	×	×	.	.	.	.	.	.
8. Miss Pearl Oglethorpe	.	.	.	.	.	×	.	×	×	.	.	.	.	.
9. Miss Ruth DeSaad	.	.	.	.	×	.	×	×	×	.	.	.	.	.
10. Miss Verne Sanderson	.	.	.	.	.	.	×	×	×	.	×	.	.	.
11. Miss Myra Liddell	.	.	.	.	.	.	.	×	×	×	×	.	.	.
12. Miss Katherine Rogers	.	.	.	.	.	.	.	×	×	×	×	×	×	×
13. Mrs. Sylvia Avondale	.	.	.	.	.	.	×	×	×	×	×	×	×	×
14. Mrs. Nora Fayette	.	.	.	.	.	×	×	.	×	×	×	×	×	×
15. Mrs. Helen Lloyd	.	.	.	.	.	.	×	×	.	×	×	×	.	.
16. Mrs. Dorothy Murchison	.	.	.	.	.	.	.	×	×	.	.	.	.	.
17. Mrs. Olivia Carleton	.	.	.	.	.	.	.	.	×	.	×	.	.	.
18. Mrs. Flora Price	.	.	.	.	.	.	.	.	×	.	×	.	.	.

Figure 5: The Southern Women data set.

The data set has been extensively studied, and used as

sw4 #1 (1,1,1,1)*(2,8,29,2)=41

Ind	G/I-Colorings	0	1	2	3	Ind	G/I-Colorings	0	1	2	3
I1	.11..311.1.31. 11111111111111	0	6	0	0	I10	...223.....3.. 22222333333333	0	0	2	2
I2	.112..11.1.3.. 11111111111111	0	5	0	0	I11	....23..2..3.. 33333333333333	0	0	0	2
I3	.112.311...31. 11111111111111	0	5	0	0	I12	....23..2..3.. 33333333333333	0	0	0	2
I4	.1.2..11.1.31. 11111111111111	0	5	0	0	I13	...223..2.23.2 22222222222222	0	0	5	0
I5	.1.2..1.....1. 11111111111111	0	3	0	0	I14	0..223.12.2..2 22222222222222	0	0	5	0
I6	.1....11...3.. 11111111133333	0	3	0	1	I15	0..22...2.23.2 22222222222222	0	0	5	0
I7	.1.2...1...3.. 33333333333333	0	0	0	1	I16	.....3.....3.. 33333333333333	0	0	0	2
I8	.....3.1...3.. 33333333333333	0	0	0	2	I17	0....3........ 00000000000000	1	0	0	0
I9	.1.2.3.....3.. 33333333333333	0	0	0	2	I18	0....3........ 00000000000000	1	0	0	0

Table 2: An optimal coloring on the Southern Women data set under cost setting $(\alpha, \beta_1, \beta_2, \gamma) = (1, 1, 1, 1)$. Cost=41

sw4 #1 (1,1,3,1)*(25,18,4,15)=70

Ind	G/I-Colorings	0	1	2	3	Ind	G/I-Colorings	0	1	2	3
I1	.11..311.1.12. 11111111111122	0	6	1	0	I10	...133.....1.. 33313333111111	0	2	0	2
I2	.111..11.1.1.. 11111111111111	0	7	0	0	I11	....33..3..1.. 33333333331111	0	1	0	3
I3	.111.311...12. 11111111111122	0	6	1	0	I12	....33..3..1.. 33333333331111	0	1	0	3
I4	.1.1..11.1.12. 11111111111122	0	6	1	0	I13	...133..3.31.3 33313333333113	0	2	0	5
I5	.1.1..1.....2. 11111112222222	0	3	1	0	I14	0..133.13.3..3 00013331333333	1	2	0	5
I6	.1....11...1.. 11111111111111	0	4	0	0	I15	0..13...3.31.3 00013333333113	1	2	0	4
I7	.1.1...1...1.. 11111111111111	0	4	0	0	I16	.....3.....1.. 33333333111111	0	1	0	1
I8	.....3.1...1.. 33333331111111	0	2	0	1	I17	0....3........ 00000000000000	1	0	0	0
I9	.1.1.3.....1.. 11111333111111	0	3	0	1	I18	0....3........ 00000000000000	1	0	0	0

Table 3: An optimal coloring on the Southern Women data set under cost setting $(\alpha, \beta_1, \beta_2, \gamma) = (1, 1, 3, 1)$. Cost=70.

a benchmark for community identification methods [13]. A summary of the community identification results of 21 methods was given by Freeman [13], and is shown as Figure 6.

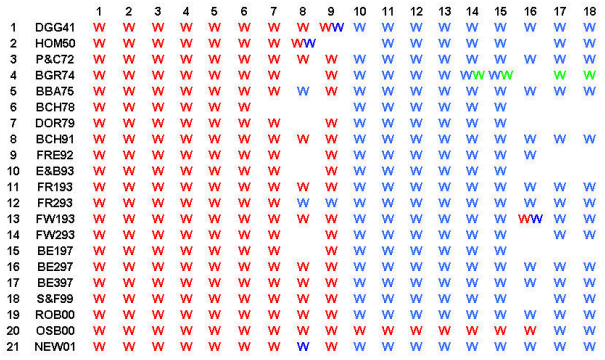


Figure 6: Communities in Southern Women according to the 21 methods in Freeman's review.

Since there is only one group present in each time step of the data set, it is essential that we use $\beta_1 \neq 0$; otherwise, the optimal solution would trivially be obtained by coloring all individuals with the same color. We considered two cost settings: $(\alpha, \beta_1, \beta_2, \gamma) = (1, 1, 1, 1)$, and $(\alpha, \beta_1, \beta_2, \gamma) = (1, 1, 3, 1)$. In the former case, temporarily switching group membership is more expensive, in the latter, being absent from a group meeting is expensive.

Tables 2 and 3 show the optimal colorings under these cost settings. The major difference between the colorings is the group color in time steps 4 and 5. Notice that our approach with both parameters correctly identifies the existence of two communities, even though only one group is present in each time step. The results of most of the 21 methods proposed in the past and our results agree that the first group consists of individuals 1–7, and the other group consists of individuals 10–13. The disagreement of the 21 results is also shown in the affiliation of individuals 8, 9, and 16. It should be noted that most of the results of previously proposed approaches can be obtained using our algorithms with appropriate parameter settings. We will elaborate on this aspect in more detail in the full version of this paper.

The coloring in Table 3 may explain why the 21 methods do not agree on the affiliation of individuals 8, 10, and 16. These individuals are affiliated with both communities for almost equal amounts of time. Furthermore, individuals 17 and 18 belong to a separate community, which explains why some of the 21 methods cannot identify their community membership.

While the greedy heuristics do not obtain optimal results on this data set, they are very close to optimal. As a representative, we show the results of the greedy algorithm using the Jaccard similarity function with the cost setting $(1, 1, 1, 1)$ in Table 4. A summary of the performance of different heuristics is given in Table 5. Due to space constraints, we defer a more in-depth discussion of the effects of heuristics to the full version of this paper.

Grevy's Zebra Data Set

The Grevy's zebra (*Equus grevyi*) data set was obtained by observing spatial proximity of members of a zebra herd. Populations of Grevy's zebras were observed by biologists [36] over three months in 2002 in Kenya. Predetermined census loops were driven approximately twice per week. Individuals were identified by unique stripe patterns, and their locations taken by GPS. In the resulting data set, individuals are in the same group if their GPS locations are very close. The data set contains 28 individuals interacting over a period of 44 time steps. Many of the individuals are missing in many time steps.

The aggregate social network of the zebra population is shown in Figure 7. Notice that the aggregate network is very dense, and would not let us infer much interesting community structure and change by itself. Figure 8 shows the result of the greedy heuristic applied to the data set. While there is no agreed-upon "correct" community structure for this relatively new data set yet, the inferred communities agree

sw4_union_52 #1 (1,1,1,1)*(0,3,41,0)=44

Ind	G/I-Colorings	0	2	4	5	6
I1	.01..200.3.20. 00000000000000	4	0	0	0	0
I2	.014..00.3.2.. 00000000000000	3	0	0	0	0
I3	.014.200...20. 00000000000000	4	0	0	0	0
I4	.0.4..00.3.20. 00000000000000	4	0	0	0	0
I5	.0.4..0.....0. 00000000000000	3	0	0	0	0
I6	.0....00...2.. 00000000000000	3	0	0	0	0
I7	.0.4...0...2.. 44444444444444	0	0	1	0	0
I8	.....2.0...2.. 22222222222222	0	2	0	0	0
I9	.0.4.2.....2.. 22222222222222	0	2	0	0	0
I10	...452.....2.. 22222222222222	0	2	0	0	0
I11	....52..5..2.. 22222222222222	0	2	0	0	0
I12	....52..5..2.. 22222222222222	0	2	0	0	0
I13	...452..5.52.5 55555555555555	0	0	0	4	0
I14	6..452.05.5..5 55555555555555	0	0	0	4	0
I15	6..45...5.52.5 55555555555555	0	0	0	4	0
I16	.....2.....2.. 22222222222222	0	2	0	0	0
I17	6....2........ 66666666666666	0	0	0	0	1
I18	6....2........ 66666666666666	0	0	0	0	1

Table 4: A heuristic coloring using Jaccard similarity on the Southern Women data set under cost setting $(\alpha, \beta_1, \beta_2, \gamma) = (1, 1, 1, 1)$. Cost=44

Cost	OPT	Jaccard	JaccardD
High i-cost	**41**	44	56
High g-cost	**70**	80	87

Table 5: Cost comparison for Southern Women

with those identified manually by biologists. Notice that the dynamic interpretation lets us observe interesting phenomena, such as individual 3 switching its affiliation during the observation period. Such changes are obscured in the static graph. Due to space constraints, a more detailed discussion of results on this data set and the performance of different heuristics is deferred to the full version.

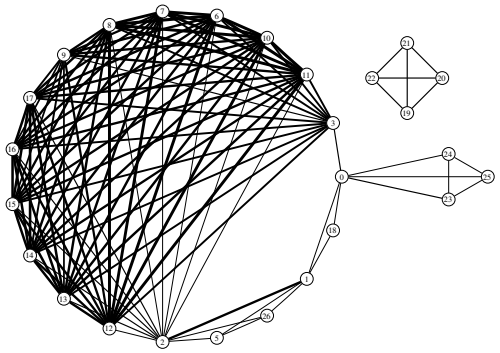


Figure 7: Aggregate social network of zebras

6. CONCLUSIONS

We have presented a framework for identifying communities and their dynamics in social networks, and shown that the results are meaningful by comparing them with traditional methods. For the Southern Women data set, our methods can identify the same communities as the traditional methods (the 21 methods in [13]). In fact, several traditional methods for community identification are special cases of the proposed framework. Moreover, our method also gives more insight into the dynamics of communities,

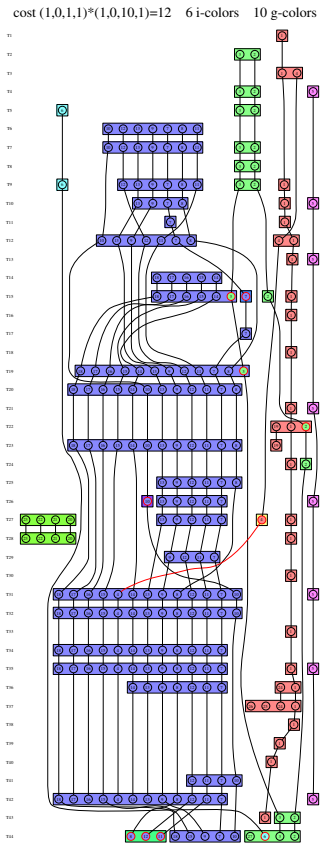


Figure 8: Coloring obtained with greedy heuristic and costs $(\alpha, \beta_1, \beta_2, \gamma) = (1, 0, 1, 1)$.

answering questions such as who changes from which community to which, and when.

Although our approach can find good communities, we showed that finding optimal colorings is NP-hard. We have presented heuristic algorithms which find near optimal solutions in practice. The next step is to devise approximation algorithms with provable guarantees in place of the heuristics used for our evaluations.

Finally, we have not investigated in full the scalability of the proposed heuristics to large networks observed over long time periods. This is an important property of all algorithms applied to current network datasets and needs to be addressed in the future.

Acknowledgments

We would like to thank Dan Rubenstein, Ilya Fischhoff, and Siva Sundaresan of the Department of Ecology and Evolutionary Biology at Princeton University for sharing the Grevy's zebra and the onager data. Their work was supported by the NSF grants CNS-025214 and IOB-9874523. We are grateful to Jared Saia for useful discussions.

7. REFERENCES

[1] J. Aizen, D. Huttenlocher, J. Kleinberg, and A. Novak. Traffic-based feedback on the web. *Proc. Nat. Acad. Sci.*, 101(Suppl.1):5254–5260, 2004.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *Proc. KDD '06*, 2006.

[3] J. Baumes, M. Goldberg, M. Magdon-Ismail, , and W. Wallace. Discovering hidden groups in communication networks. *Proc. 2nd NSF/NIJ Symp. on Intelligence and Security Informatics*, 2004.

[4] T. Y. Berger-Wolf and J. Saia. A framework for analysis of dynamic social networks. In *Proc. KDD '06*, 523–528, 2006.

[5] K. Carley, M. Prietula, and editors. *Computational Organization Theory*. Lawrence Erlbaum associates, Hillsdale, NJ, 2001.

[6] D. P. Croft, R. James, P. Thomas, C. Hathaway, D. Mawdsley, K. Laland, and J. Krause. Social structure and co-operative interactions in a wild population of guppies (*poecilia reticulata*). *Behavioural Ecology and Sociobiology*, In Press.

[7] P. C. Cross, J. O. Lloyd-Smith, , and W. M. Getz. Disentangling association patterns in fission-fusion societies using African buffalo as an exa. *Animal Behaviour*, 69:499–506, 2005.

[8] E. Dahlhaus, D. Johnson, C. Papadimitriou, P. Seymour, and M. Yannakakis. The complexity of multiterminal cuts. *SIAM J. Comput.*, 23:864–894, 1994.

[9] A. Davis, B. B. Gardner, and M. R. Gardner. *Deep South*. The U. of Chicago Press, Chicago, IL, 1941.

[10] I. R. Fischhoff, S. R. Sundaresan, J. Cordingley, H. M. Larkin, M.-J. Sellier, and D. I. Rubenstein. Social relationships and reproductive state influence leadership roles in movements of plains zebra (*equus burchellii*). *Animal Behaviour*, 2006. Submitted.

[11] G. Flake, S. Lawrence, C. L. Giles, and F. Coetzee. Self-organization and identification of web communities. *IEEE Computer*, 35(3), 2002.

[12] L. Freeman. On the sociological concept of "group": a empirical test of two models. *American Journal of Sociology*, 98:152–166, 1993.

[13] L. Freeman. Finding social groups: A meta-analysis of the southern women data. In R. Breiger, K. Carley, and P. Pattison, editors, *Dynamic Social Network Modeling and Analysis*. The National Academies Press, Washington, D.C., 2003.

[14] L. Getoor and C. P. Diehl. Link mining: a survey. *SIGKDD Explor. Newsl.*, 7(2):3–12, 2005.

[15] D. Gibson, J. M. Kleinberg, and P. Raghavan. Inferring web communities from link topology. In *UK Conference on Hypertext*, 225–234, 1998.

[16] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci.*, 99:8271–8276, 2002.

[17] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proc. WWW '04*, 491–501, 2004.

[18] J. Hopcroft, O. Khan, B. Kulis, and B. Selman. Natural communities in large linked networks. In *Proc. KDD '03*, 541–546, 2003.

[19] P. Jaccard. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:241–272, 1901.

[20] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. 2003.

[21] M. Kretzschmar and M. Morris. Measures of concurrency in networks and the spread of infectious disease. *Math. Biosci.*, 133:165–195, 1996.

[22] R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. Trawling the web for emerging cyber-communities. In *Proc. WWW '99*, 1999

[23] M. Magdon-Ismail, M. Goldberg, W. Wallace, and D. Siebecker. Locating hidden groups in communication networks using hidden markov models. *Proc. ISI '03*, 2003.

[24] B. Malin. Data and collocation surveillance through location access patterns. *Proc. NAACSOS Conf.*, 2004.

[25] L. A. Meyers, M. Newman, and B. Pourbohloul. Predicting epidemics on directed contact networks. *Journal of Theoretical Biology*, 240:400–418, 2006.

[26] M. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev.*, 69, 2004.

[27] M. Newman, A.-L. Barabási, and D. J. Watts, editors. *The Structure and Dynamics of Networks*. Princeton University Press, 2006.

[28] J. M. Read and M. J. Keeling. Disease evolution on networks: the role of contact structure. *Proc. R. Soc. Lond. B*, 270:699–708, 2003.

[29] E. M. Rogers. *Diffusion of Innovations*. Simon & Shuster, Inc., 5th edition, 2003.

[30] D. I. Rubenstein, S. Sundaresan, I. Fischhoff, and D. Saltz. Social networks in wild asses: Comparing patterns and processes among populations. In A. Stubbe, P. Kaczensky, R. Samjaa, K. Wesche, and M. Stubbe, editors, *Exploration into the Biological Resources of Mongolia*, volume 10. Martin-Luther University Halle-Wittenberg, 2007. In press.

[31] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. In *Molecular biology and evolution*, 1987.

[32] D. Sedley. The stoic criterion of identity. *Phronesis*, 27:255–275, 1982.

[33] K. L. Spalding, R. D. Bhardwaj, B. A. Buchholz, H. Druid, and J. Frisén. Retrospective birth dating of cells in humans. *Cell*, 122(1):133–143, 2005.

[34] G. Simmel. *The Sociology of Georg Simmel*. ed. by K.H. Wolff. Free Press, Glencoe, IL, 1950.

[35] G. Simmel. *Conflict and the Web of Group Affiliations*. Free Press, Glencoe, IL, 1955.

[36] S. R. Sundaresan, I. R. Fischhoff, J. Dushoff, and D. I. Rubenstein. Network metrics reveal differences in social organization between two fission-fusion species, Grevy's zebra and onager. *Oecologia*, 2006.

[37] T. Uno. Algorithms for enumerating all perfect, maximum and maximal matchings in bipartite graphs. In *ISAAC '97: Proceedings of the 8th International Symposium on Algorithms and Computation*, pages 92–101, London, UK, 1997. Springer-Verlag.

[38] S. Wasserman and K. Faust. *Social Network Analysis*. Cambridge University Press, Cambridge, MA, 1994.