A Framework for Analysis of Dynamic Social Networks

Tanya Y. Berger-Wolf*
Department of Computer Science
University of Illinois at Chicago
851 S. Morgan (M/C 152)
Chicago, IL 60304
tanyabw@cs.uic.edu

Jared Saia†
Department of Computer Science
University of New Mexico
Albuquerque, NM 87131
saia@cs.unm.edu

ABSTRACT

Finding patterns of social interaction within a population has wide-ranging applications including: disease modeling, cultural and information transmission, and behavioral ecology. Social interactions are often modeled with networks. A key characteristic of social interactions is their continual change. However, most past analyses of social networks are essentially static in that all information about the time that social interactions take place is discarded. In this paper, we propose a new mathematical and computational framework that enables analysis of dynamic social networks and that explicitly makes use of information about when social interactions occur.

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

General Terms: Algorithms, Design

Keywords: dynamic social networks, algorithms, disease spread.

1. INTRODUCTION

Finding patterns of social interaction within a population has wide-ranging applications including: disease modeling [14, 22], cultural and information transmission [4, 6, 9, 18, 30, 32], intelligence and surveillance [4, 21, 26], business management [5, 8, 28, 29], conservation biology and behavioral ecology [10, 11, 25]. The most common way to capture information on social interactions is a network [15, 16, 20, 27, 31].

Typically, each individual is represented by a node in the network, and there is an edge between two nodes if a social interaction has occurred at any point in time between the two individuals represented by these nodes. Depending on the source of data, a social interaction could be a verbal or written communication (cellphones, emails, blogs, chatrooms, etc.), scientific collaboration (co-authorship networks), sexual contact (HIV patients, dating among adolescents), or physical or virtual proximity (visiting websites, physical locations, groups of animals). Edges are commonly weighted by frequency of interaction.

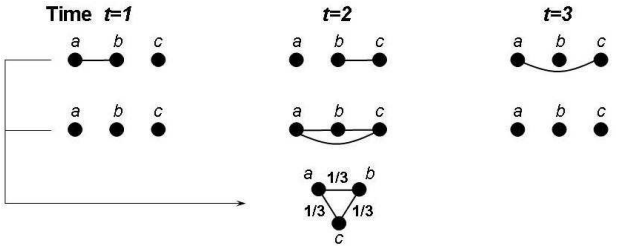

Figure 1: The top two rows are two dynamic graphs that map to the same static graph (third row).

This network model of social interactions has been very successfully. However, a major drawback of this model is that it is essentially static in that all information about the time that social interactions take place is discarded. The static nature of the model can give inaccurate or inexact information about patterns in the data. For example, as Figure 1 illustrates, in some cases, very different dynamic data can give rise to the same static graph. Thus, decisions made based solely on the static data may be flawed. For example, assume the edges in Figure 1 represent social contact that can cause disease transmission. Suppose we can vaccinate only one person and we want to minimize the total number of infected individuals at the end, assuming that the disease always crosses each edge and that any individual may be infected initially. From the static graph, it seems that no matter which single individual is vaccinated there will be two individuals infected at the end. However, in the first dynamic graph, assuming the initial infection starts at step 0, only b needs to be vaccinated so that no matter whether a or c is infected initially, the infection will not spread.

The static graph representation prevents us from even asking certain fundamental questions about either the causes or consequences of social patterns. How quickly can a disease spread through the population and which individuals should we inoculate to slow down its spread? How do the

*Was supported by the National Science Foundation postdoctoral fellowship grant EIA 02-05116

†Supported by the National Science Foundation grant CCR-0313160 and by the Sandia University Research Program Grant No. 191445

KDD'06, August 20–23, 2006, Philadelphia, Pennsylvania, USA.

size and stability of social structures change with outside circumstances (e.g. season, time of day, predator activity, upcoming conference or journal deadlines, court subpoenas, terrorist activities)? Are there differences in the life span of social structures with respect to their size and the demographics of their members? To be able to answer these questions, we need to have information on when social interactions occurred. This level of information is becoming increasingly available, but the analytical and computational tools are still lagging.

Research in dynamic network analysis has proceeded in several directions. The statistical mechanics view [2, 3] considers networks as complex physical systems and strives to describe laws governing their evolution and limit behavior and properties. A more computational view [7] incorporates probabilities and uncertainty into the structure information and combines social network analysis with multi-agent systems. Computer simulations until recently have been the main computational technique to incorporate dynamic network information, e.g. [14]. The last few years have seen a development of systematic algorithmic approaches to dynamic network analysis, mostly in the context of information networks [1, 18, 20, 19, 23, 24]. Yet, most of the methods focus on the *frequency*, rather than *concurrency* and *order* of interactions. Moreover, most dynamic models are network evolution models where nodes and edges can be added but not deleted over time.

In this paper, we propose a new mathematical and computational framework that enables analysis of dynamic social networks and that explicitly makes use of information about the time that social interactions occur. We present several algorithms for obtaining information about the structure of dynamic social networks in this framework and demonstrate the utility of these algorithms on real data.

The rest of the paper is organized as follows: in Section 2 we describe our new framework for analyzing dynamic social networks. In Section 3 we present algorithms that use this framework to find some basic properties of dynamic social networks. In particular, we give algorithms for finding the most persistent and the largest social structures, as well as social structures that encompass a set of specified groups of individuals. Section 6 lists a collection of open problems and comments on some possible solution approaches.

2. OUR FORMAL MODEL

Input Data: To answer questions about social structure in a dynamic setting, we assume that a population of individuals is monitored in some way over a period of time. Interactions between individuals are recorded at every timestep. For the purpose of this paper, we assume that the input is in the form of the partition of the individuals into groups at every time step. Given a population $X = \{x_1, ..., x_n\}$, we define a *group* to be a subset $g \subseteq X$. We assume that the input is a set of partitions, $P_1, P_2, ..., P_T$ of X, one partition for each time step. Each partition, P_i, is a set of disjoint *groups*. We denote by $P(g)$ the index of the partition to which g belongs. That is, if $g \in P_i$ then $P(g) = i$.

Given two groups, g and h, a set similarity measure $sim(,)$, and a *turnover threshold* β, the two groups are *similar* if $sim(g, h) \geq \beta$. Our definition is independent of any specific set similarity measure. There are many possible such measures, including the standard Jaccard similarity measure [17] (the size of the intersection over the union). We do assume that the similarity measure is efficiently computable and we also assume the following properties hold:

1. $sim(g, g)$ has maximum similarity value.
2. $sim(h, g)$ monotonically increases with the increase of $|h \cap g|$ when $|h| + |g|$ is fixed.
3. $sim(h, g)$ monotonically increases with the decrease of $|h| + |g|$ when $|h \cap g|$ is fixed.

We now define the main concept of our framework - a metagroup.

DEFINITION 1. *Given partitions $P_1, ..., P_T$ of a set of individuals X, a set similarity measure $sim(,)$, a turnover threshold β and a function $\alpha(T)$, a* metagroup *MG is a sequence of groups $MG = \langle g_1, ..., g_l \rangle$, $\alpha(T) \leq l \leq T$ such that*

1. *no two groups in MG are in the same partition and the groups are ordered by the partition time steps:*

$$\forall i, j, \quad 1 \leq i < j \leq l, \quad P(g_i) < P(g_j),$$

2. *the consecutive groups in MG are "similar" in that:*

$$\forall i, \quad 1 \leq i < l, \quad sim(g_i, g_{i+1}) \geq \beta.$$

We call the parameter α the persistence *of a metagroup.*

Note, that the intersection between g_1 and g_l may be null by this definition; our only constraint is that the groups change gradually (as defined by β).

DEFINITION 2. *An individual $x \in X$ is a* member of a metagroup *$MG = \langle g_1, ..., g_l \rangle$ if the number of groups $g_1, ..., g_l$ to which x belongs is at least an a priori chosen* membership threshold *function γ (which may be a function of T, the total number of individuals associated with MG, and other parameters).*

The values of α (persistence), β (turnover) and γ (membership), give the meaning of a "group". Our framework is independent of these definitions and is capable of providing significant answers for a wide range of applications.

We use a weighted multipartite directed graph for the conceptual representation: $G = (V_1, ..., V_T, E)$ where V_i is the set of groups in partition P_i and $(g_i, g_j) \in E$ if $P(g_i) < P(g_j)$ and $sim(g_i, g_j) \geq \beta$. Note that this is a directed acyclic graph (DAG) since all the edges are directed from an earlier time step to a later one. The weight $w(g_i, g_j) = sim(g_i, g_j)$. A metagroup in this graph is a path of length at least $\alpha(T)$. We shall call this graph a *metagroup β-graph*.

From now on we assume the definition of a metagroup that satisfies a priori given thresholds of α and β and an individual membership that satisfies a threshold γ. We assume the input is in the form of a metagroup β-graph described above with the present edges of weight no less than β.

3. BASIC ALGORITHMS

Metagroup Statistics: It may be impractical to list all the metagroups; in some cases there are an exponential number. Nonetheless, we can calculate many statistics efficiently, such as the number of metagroups and the average metagroup lifespan. The number of metagroups is the number of paths of length at least α in the β-graph and the average

metagroup length is the average length of a path that is at least α long. Both can be computed with the same simple dynamic programming algorithm presented below. Let $P(g,l)$ denote the number of paths of length exactly l from a minimal vertex (no incoming edges) to a group g. We compute the table $P(g,l)$ starting with the groups in partition 0 and moving forward in time to partition T.

Alg METAGROUPS STATISTICS
$\forall g, 0 \le l \le P(g)$

$$P(g,l) = \begin{cases} 1 & \text{if } g \text{ is minimal and } l = 0 \\ 0 & \text{if } g \text{ is minimal and } l > 0 \\ \sum_{(h,g)\in E} P(h, l-1) & \text{otherwise} \end{cases}$$

Total number of metagroups: $N(MG) = \sum_{\substack{\text{maximal } g \\ l \ge \alpha}} P(g,l)$

Average metagroup length: $AL(MG) = \frac{\sum_{\substack{\text{maximal } g \\ l \ge \alpha}} P(g,l) \times l}{N(MG)}$

Maximal metagroup length: $MaxL(MG) = \max_{\substack{\text{maximal } g \\ l \ge \alpha, P(g,l) > 0}} \{l\}$

Extremal Metagroups: A fundamental question about social groups is that of their persistence. MOST Persistent Metagroup: *Find a metagroup MG which maximizes the number of groups associated with MG.* The question is equivalent to finding the longest path in a DAG, which is a well studied problem and can be solved in linear time using dynamic programming on a topologically sorted graph.

Alternatively, one can ask for the most stable (least turnover) metagroup. In our framework this is equivalent to the path of length at least α with the heaviest average edges. Most Stable Metagroup: *Find a metagroup MG with the maximum sum of the edge weights divided by the length of the path.* Again, in a DAG, such a path can be easily found using dynamic programming on a topologically sorted graph.

We can also ask for a metagroup with the largest membership. LARGEST Metagroup: *Find the metagroup MG which maximizes the number of members of MG.* The following simple algorithm solves this problem.

Alg LARGEST METAGROUP
Initialize:

$\forall g$, s.t. $P(g) = 0$, $S(g,k,l) = \begin{cases} g & \text{if } k = 1, l = 0 \\ \emptyset & \text{otherwise} \end{cases}$
$\forall g$, $S(g,0,0) = X$
$\forall g, \forall l > 0$, $S(g,0,l) = S(h_{\max}, 0, l-1)$,
where $|S(h_{\max}, 0, l-1)| = \max_{(h,g)\in E} |S(h,0,l-1)|$
$N(g,0,l) = h_{\max}$

Fill the table:
$\forall g, \forall l > 0, k > 0$,
$S(g,k,l) = (S(h_{\max}, k, l-1) - g) \cup (S(h_{\max}, k-1, l-1) \cap g)$
where $h_{\max}$ achieves
$\max_{(h,g)\in E} |S(h,k,l-1) - g| + |S(h,k-1,l-1) \cap g|$
$N(g,k,l) = h_{\max}$

Construct the metagroup:
Let g_{max} be a group s.t. $|S(g_{max}, \gamma, \alpha)| = \max_g |S(g,\gamma,\alpha)|$
Trace back the corresponding metagroup MG using the $N(g_{max}, \gamma, \alpha)$ entry as a starting point.
Return MG.

In this algorithm, $S(g,k,l)$ is the *set* of individuals that appear at least k times in a metagroup of length at least l that ends at the current group g. $N(g,k,l)$ keeps track of the incoming edge of the metagroup to which g is assigned. ($P(g)$ is the index of the partition to which g belongs). The worst case running time of the algorithm is $O(mT^2n|E|)$, where m is the total number of groups.

NAMES OF PARTICIPANTS IN GROUP I	(1) 6/27	(2) 2/2	(3) 4/12	(4) 9/26	(5) 2/25	(6) 5/19	(7) 2/15	(8) 9/16	(9) 4/8	(10) 6/10	(11) 2/23	(12) 4/7	(13) 11/21	(14) 5/3
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 DeSand					×		×	×	×					
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 2: The Southern women data

Note that all of the various metagroup statistics we have discussed so far can be computed in one pass over the data, while building the metagroups graph.

4. EXAMPLE ANALYSIS

We demonstrate the conceptual metagroup framework on a dataset which is considered a benchmark in comparing social networks analysis methods [15], the Southern women data from 1930s Natchez, Mississippi [12]. The dataset has been used to compare methods that identify communities from social interactions and the core versus the periphery members within each community [15]. Figure 2 shows the data of participation of the 18 women in 14 social activities over the nine month period [12]. Figure 3 shows the corresponding metagroups graph for $\beta = .6$. Each vertex represents a group defined by a social event. The groups are ordered chronologically and numbered using the numbering from the table in Figure 2. The similarity between any two groups g and h is computed using the generalized Jaccard measure $sim(g,h) = \frac{2|g \cap h|}{|g|+|h|}$. Only the edges of weight .6 or greater are shown. The metagroup graph has three con-

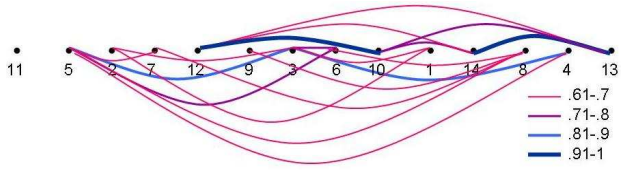

Figure 3: The metagroup graph for the Southern women data with $\beta = .6$

nected components:

- $\{11\}$, an event not strongly similar to any other event.
- $MG1 = (12, 10, 14, 13)$ is a set of events whose membership, depending on the value of γ is as follows:
  $\gamma = 1$: (the union) 10, 11, 12, 13, 14, 15
  $\gamma = 2$: 11, 12, 13, 14, 15
  $\gamma = 3, 4$: (the intersection) 12, 13, 14
- $\{5, 2, 7, 9, 3, 6, 1, 8, 4\}$ is a set of events with much more fluid membership. The longest metagroup in this component is $MG2 = (5, 3, 6, 8)$ with membership ranging from $1 - 16$ for $\gamma = 1$ (union) to $\{1 - 4, 6\}$ for $\gamma = 4$ (intersection). The most stable metagroup is

$MG3 = (5, 3, 4)$ with membership from $\{1-7, 9\}$ for $\gamma = 1$ to $\{1, 3, 4, 5\}$ for $\gamma = 3$.

Freeman [15] compares 21 different social network analysis techniques on this dataset. Figure 4 shows the assignment of individuals to clusters by different methods. Using the most stable metagroups $MG1$ and $MG3$, our method separates the individuals into clusters (using $\gamma = 1$) $\{1-7, 9\}$ and $\{10-15\}$, which is exactly as the algebraic topology based DOR79 method [13] with competence score of .923 [15]. In

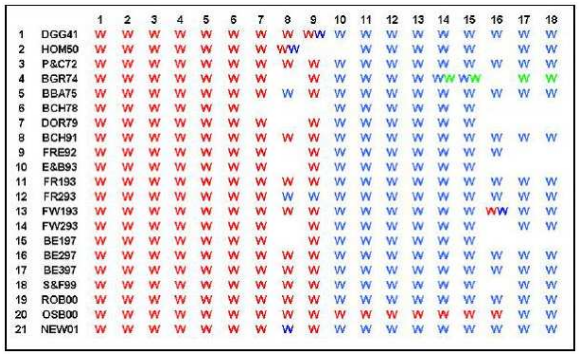

Figure 4: Clusters assigned by 21 procedures. Clusters are designated by colors. An individual assigned to more than one cluster has "W"s of two colors.

addition, Freeman compares the assignment of individuals to the core and the periphery of a cluster by different methods. Figure 5 summarizes the results. Our method (using the range of γ) separates the first cluster as 1 3 4 5|2 6|7 9 and the second as 12 13 14|11 15|10, which is different from DOR79 but comparable to the overall trend. Overall, our method provides insights comparable to the other techniques. However, the main advantage of our approach is that it explicitly addresses the time component of the data and provides a unifying framework for a more sophisticated analysis, as demonstrated in the following section.

5. CRITICAL NETWORK PROPERTIES

Group Connectivity: A natural question to ask is, given a collection of groups at different time steps, do they belong

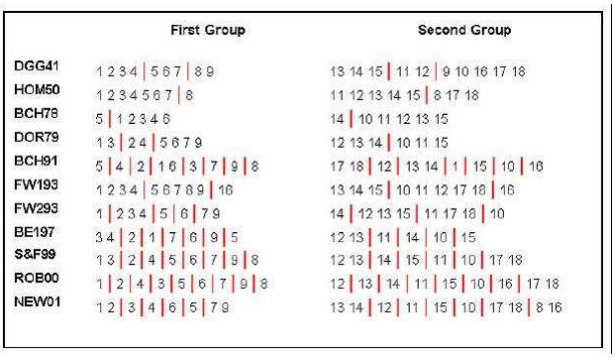

Figure 5: Core/Periphery assignment by 11 procedures. The bars separate the groups in decreasing order of "centrality".

to the same overall social structure, i.e. the same metagroup. Formally, there are several computational problems of GROUP CONNECTIVITY: Let $g_1, ..., g_l$ be a set of groups in separate partitions, ordered by their partition indices (i.e., $P(g_i) < P(g_{i+1})$). Then we may ask (1) Is there a metagroup that contains all the groups? (2) Find the most persistent/stable/largest metagroup that contains all groups. (3) If no metagroup contains all the groups, find the one that contains most. To answer all of these questions we first preprocess the graph (using simple BFS) to find a connected component with the vertices $g_1, ..., g_l$ being the cut vertices and all the minimal vertices being in partitions before g_1 (or g_1 itself) and all maximal vertices being in partitions after g_l (or g_l itself). We shall call this subgraph a *β-component* of $g_1, ..., g_l$. With this β-component graph we can answer the various group connectivity questions.

Is there a metagroup that contains all the groups? If the β-component of the groups is connected then the question is whether there is a path from a minimal to a maximal vertex of length at least α.

Find the most persistent/stable/largest metagroup that contains all groups. This is equivalent to finding the longest/average heaviest/largest membership path from a minimal to a maximal vertex in a connected β-component. We return such a path if it is of length at least α.

Find the metagroup that contains the maximum number of groups $\{g_1, ..., g_l\}$. A dynamic programming algorithm that tracks $M(g)$, the maximum number of groups $\{g_1, ..., g_l\}$ in a metagroup that ends at g.

Individual Connectivity: Similar to group connectivity, we may ask questions about individual connectivity. That is, given a set of individuals, do they belong to the same social structure. More formally, INDIVIDUAL CONNECTIVITY: Let $S \subseteq X$ be a set of individuals.

Find the metagroup that contains the largest number of individuals in S as members. This is equivalent to the LARGEST METAGROUP problem over S on the originial metagroup β-graph.

Find the most persistent metagroup that contains all the individuals in S as its members. Again, we use the LARGEST METAGROUP algorithm with slight modifications.

Find the largest metagroup that contains all the individuals in S as members. Again, a modification of the LARGEST METAGROUP answers this question.

Critical Group Set: Until now, we have been able to solve all of our problems in polynomial time. However, it would be naive to expect that all the aspects of a dynamic social network can be explored so efficiently. For example, the question of fragility of a social network can be formulated in many ways. One way to ask it is to consider the smallest set of groups whose absence would leave no recognizable overall social structures. Formally, CRITICAL GROUPS SET: Find the smallest set of groups whose removal leaves no metagroups (with respect to given β and α). This question is particularly important in an epidemiological context where it is safe to have prolonged interactions for periods no longer than the non-contagious incubation period of a disease, yet it is important to quarantine people otherwise. In this context some groups may be considered more important (or easier) to dissipate than others and this would lead to a vertex-weighted version of the problem. Alternatively, these are the events where vaccination stations should be positioned for mass vaccination. For example, flu vaccines on

university campuses may be administered more effectively outside of some lecture halls than others. More generally, MIN K-PATH VERTEX SHATTERING SET: for an arbitrary graph $G = (V, E)$ find the smallest (weighted) subset of vertices $U \subseteq V$ such that the subgraph induced by $V - U$ has no paths longer than $k - 1$. We show that the complexity of this problem very much depends on the specific value of k.

First, we show that the general problem is NP-complete by demonstrating that the unweighted MIN 2-PATH VERTEX SHATTERING SET is equivalent to MIN VERTEX COVER, which is a well studied NP-complete problem.

THEOREM 1. *MIN 2-PATH VERTEX SHATTERING SET is polynomially equivalent to* MIN VERTEX COVER.

COROLLARY 1. *MIN K-PATH VERTEX SHATTERING SET on an arbitrary graph G is NP-complete.*

The special structure of a metagroup β-graph does not improve the complexity of the MIN 2-PATH VERTEX SHATTERING SET problem.

THEOREM 2. *MIN 2-PATH VERTEX SHATTERING SET on a metagroup β-graph is NP-complete.*

While the MIN k-PATH VERTEX SHATTERING SET problem is NP-hard for an arbitrary k, we present a polynomial time algorithm for $k = T$ in a DAG, where T is the length of the longest path in G. Let $P(v, l)$ be the number of paths of length exactly l from a minimal vertex to $v \in V$. The algorithm uses the function K-GRAPH:

Alg K-GRAPH(k)
In the order of a BFS from the minimal vertices: $\forall v, l, \quad 0 \leq l \leq T$

$$P(v,l) = \begin{cases} 1 & \text{if } v \text{ is minimal and } l = 0 \\ 0 & \text{if } v \text{ is minimal and } l > 0 \\ \sum_{(u,v)\in E} P(u, l-1) & \text{otherwise} \end{cases}$$

FOR all maximal v and $l \geq k$ s.t. $P(v, l) > 0$ **DO**
Trace back the paths of length l that terminate at v
RETURN the resulting traced graph G_k

LEMMA 1. *Algorithm k-GRAPH(k) returns a subgraph G_k of G that contains every edge in G which is in some path of length at least k from a minimal to a maximal vertex in G.*

COROLLARY 2. *G_k contains all the paths of length at least k from a minimal to a maximal vertex in G.*

The running time of the K-GRAPH algorithm is $O(|E||V|)$ since each edge is traced exactly once and $T \leq |V|$. Using the function K-GRAPH we can state the MIN T-PATH VERTEX SHATTERING SET solution.

Alg MIN T-PATH VERTEX SHATTERING SET G_T = K-GRAPH(T)
Add a vertex s connected to all the minimal vertices and a vertex t connected from all the maximal vertices in G_T
U = minimal vertex cut in the resulting graph $G_T + \{s, t\}$
RETURN U.

THEOREM 3. *Algorithm* MIN T-PATH VERTEX SHATTERING SET *is a polynomial time algorithm that returns the smallest subset of vertices U whose removal leaves a subgraph with no path of length T (where T is the length of the longest path in the original graph G).*

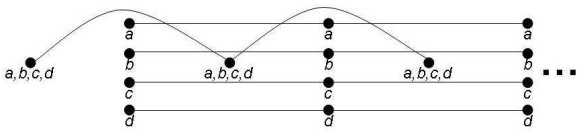

Figure 6: Here two groups are similar only if they are identical. The removal of any individual does not change the structure of the graph until there is only one individual left. Until that point the singleton paths are $T/2$ long and do not intersect the path connecting the group of all the individuals. When there is only one individual left then the one path remaining is T long.

Critical Individual Set: Another way to address the question of the fragility of a social network is to ask what is the smallest number of *individuals* whose absence would leave no recognizable social structures. In the epidemiological setting, for example, vaccination of those individuals would prevent the spread of a disease (with the right definitions of a group and a metagroup). CRITICAL INDIVIDUALS SET: Find the smallest set of individuals whose removal leaves no metagroups.

First we note that removing individuals from the population does not guarantee that the metagroups become less persistent (the paths in the metagroups graph become shorter). In fact, Figure 6 shows an example of a metagroups graph for which the removal of all but one individual (no matter which one) leads to a doubling of the path length. While the example in Figure 6 may seem contrived, in fact the same phenomenon occurs in *every* metagroups β-graph where β is such that only identical groups are connected.

The example above shows that the identity similarity measure may be impractical, particularly in the context of disease or information spreading. The identity similarity measure assumes that even groups that share most of their members but are not identical do not pass information between them or transmit diseases. Once we relax the requirement for the two groups to be identical, both the realistic implications and the combinatorics of the critical individual set change drastically.

PROPOSITION 1. *For any edge in the graph (h, g) and any individual x there are three possible ways in which the removal of x can affect the weight of the edge (h, g):*

1. *If $x \notin h$ and $x \notin g$ then $sim(h, g)$ does not change with the removal of x.*
2. *If $x \in h \cap g$ then $sim(h, g)$ will not increase with the removal of x, but may decrease.*
3. *If $x \in h$ but $x \notin g$ (or vice versa) then $sim(h, g)$ will not decrease with the removal of x, but may increase.*

The proof of this is straight forward once we recall the three properties of the similarity measure.

There are several greedy heuristics for this problem. In particular, one may iteratively remove the individual that: 1) appears in the intersections of the largest number of groups that are still connected by an edge; 2) removes the largest number or the heaviest of edges; 3) reduces the (total) weight of the edges by most; or 4) removes edges from the largest number of metagroups.

6. EXTENSIONS AND OPEN PROBLEMS

There are many other questions one may ask about a dynamic network that can be easily modeled computationally in our framework. We give a list of open problems here.

Individual Membership: Given an individual x, find the metagroup MG which maximizes the cardinality of the set of groups in MG in which x occurs. That is, $|\{i,\ s.t.\ x \in g_i,\ g_i \in MG\}|$

$$= \max_{\text{metagroups } D} |\{j,\ s.t.\ x \in g_j,\ g_j \in D\}|.$$

Extroverts and Introverts: Find the individual who is a member of the largest (smallest) number of metagroups.

Loyal Individuals: Given an individual, what fraction of its time is it a member of the same metagroup? Find the individuals that appear most frequently in one metagroup.

Metagroup Representative: Given a metagroup MG, is there an individual who occurs more in this metagroup than any other individual and occurs in MG more than in any other metagroup?

Demographic Distinction: Given a coloring of individuals (a partition), is there a property that distinguishes one color from the others, i.e. some color is in more metagroups, fewer metagroups, longer metagroups, more time in any metagroup (e.g. on average), etc.? (Each color represents a demographic set.)

Critical Parameter Values: Identify the largest values of α, β for which there exists at least 1 (k) metagroup. Identify the largest value of γ for which each metagroup has at least one member.

Critical Time Moments: Identify critical time moments. For example, the time when groups' membership changes most, i.e. minimal edge cut.

For all of the above questions it is important also to keep in mind that the input graph may be very large and even polynomial algorithms may be too slow. Thus, the real goal in this domain would be to design sublinear algorithms.

Acknowledgments

We are grateful to S. (Muthu) Muthukrishnan, Dan Rubenstein, Siva Sundaresan, Ilya Fischoff, David Kempe, Simon Levin and many others for their help and insights.

7. REFERENCES

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

[2] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74:47–97, 2002.

[3] A.-L. Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207–211, 2005.

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

[5] S. Bernstein, A. Clearwater, S. Hill, C. Perlich, and F. Provost. Discovering knowledge from relational data extracted from business news. *Proc. Workshop on Multi-Relational Data Mining*, 2002

[6] K. Carley. Communicating new ideas: The potential impact of information and telecommunication technology. *Tech. in Society*, 18(2):219–230, 1996.

[7] K. Carley. Dynamic network analysis. In R. Breiger, K. Carley, and P. Pattison, eds, *Dynamic Social Network Modeling and Analysis*, 133–145. The Nat. Acad. Press, Wash., DC, 2003.

[8] K. Carley and M. Prietula, eds. *Computational Organization Theory*. Lawrence Erlbaum ass., Hillsdale, NJ, 2001.

[9] L. Chen and K. Carley. The impact of social networks in the propagation of computer viruses and countermeasures. *IEEE Trans. Sys., Man and Cyber.*, forthcoming.

[10] D. Croft, J. Krause, and R. James. Social networks in the guppy (*Poecilia Reticulata*). *Proc. Royal Soc. London. Series B. Bio. Sci*, 271:516–519, 2004.

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

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

[13] P. Doreian. On the delineation of small group structure. In H. C. Hudson, ed, *Classifying Social Data*. Jossey-Bass, San Francisco, CA, 1979.

[14] S. Eubank, H. Guclu, V. Kumar, M. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429:429:180–184., Nov 2004. Supplement material.

[15] 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 Natl Acad. Press, Washington, D.C., 2003.

[16] M. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[17] P. Jaccard. The distribution of flora in the alpine zone. *The New Phytologist*, 11(2):37–50, 1912.

[18] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. *Proc. 9th ACM SIGKDD*, 2003.

[19] J. Kleinberg. Temporal dynamics of on-line information streams. Draft chapter for the forthcoming book Data Stream Management: Processing High-Speed Data Streams (M. Garofalakis, J. Gehrke, R. Rastogi, eds.), Springer.

[20] J. Kleinberg. Small-world phenomena and the dynamics of information. In *Proc. 17th Intl. Joint Conf. Artif. Intel.*,, Morgan Kaufman, 2001.

[21] G. Kolata. Ideas and trends; Enron offers an unlikely boost to e-mail surveillance. *New York Times*, May 22 2005.

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

[23] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. *Proc. Intl WWW Conf.*, 2003.

[24] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: Densification laws, shrinking diameters and possible explanations. *Proc. 11th ACM SIGKDD*, 2005.

[25] D. Lusseau and M. E. J. Newman. Identifying the role that individual animals play in their social network. *Proc. R. Soc. London B (Suppl.)*, 271:S477–S481, 2004.

[26] M. Magdon-Ismail, M. Goldberg, W. Wallace, and D. Siebecker. Locating hidden groups in communication networks using hidden markov models. In *Proc. Intl. Conf. on Intel. and Security Inform.*, 2003.

[27] M. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[28] C. Papadimitriou. Computational aspects of organization theory. *Lecture Notes in Computer Science*, 1997.

[29] C. Papadimitriou and E. Servan-Schreiber. he origins of the deadline: Optimizing communication in organizations. *Complexity in Economics.*, 1999.

[30] J. Tyler, D. Wilkinson, and B. Huberman. Email as spectroscopy: Automated discovery of community structure within organizations. *Proc. 1st Intl. Conf. on Comm. and Tech.*, 2003.

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

[32] B. Wellman. An electronic group is virtually a social network. In S. Kiesler, editor, *Culture of the Internet*, pages 179–205. Lawrence Erlbaum, Mahwah, NJ, 1997.