AFRL-RI-RS-TM-2012-001

EVOLVE: ANALYZING EVOLVING SOCIAL NETWORKS

FETCH TECHNOLOGIES

JULY 2012

FINAL TECHNICAL MEMORANDUM

APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED

STINFO COPY

AIR FORCE RESEARCH LABORATORY
INFORMATION DIRECTORATE

■ **AIR FORCE MATERIEL COMMAND** ■ **UNITED STATES AIR FORCE** ■ **ROME, NY 13441**

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th ABW, Wright-Patterson AFB Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RI-RS-TM-2012-001 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

FOR THE DIRECTOR:

/ S /

TODD V. WASKIEWICZ
Work Unit Manager

/ S /

MICHAEL J. WESSING
Deputy Chief, Information Intelligence
Systems and Analysis Division
Information Directorate

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Washington Headquarters Service, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC 20503.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
JULY 2012	FINAL TECHNICAL MEMORANDUM	MAR 2011 – MAY 2012

4. TITLE AND SUBTITLE
EVOLVE: ANALYZING EVOLVING SOCIAL NETWORKS

5a. CONTRACT NUMBER
FA8750-11-C-0127

5b. GRANT NUMBER
N/A

5c. PROGRAM ELEMENT NUMBER
63788F

6. AUTHOR(S)
Sofus Macskassy

5d. PROJECT NUMBER
E3NA

5e. TASK NUMBER
DS

5f. WORK UNIT NUMBER
NA

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

PRIME:
Fetch Technologies
841 Apollo St, Suite 400
El Segundo, CA 90245

SUB:
Information Sciences Institute
University of Southern California
4676 Admiralty Way
Marina del Ray, CA 90292

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Research Laboratory/RIEA
525 Brooks Road
Rome NY 13441-4505

10. SPONSOR/MONITOR'S ACRONYM(S)
N/A

11. SPONSORING/MONITORING AGENCY REPORT NUMBER
AFRL-RI-RS-TM-2012-001

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited. PA# 88ABW-2012-3677
Date Cleared:

13. SUPPLEMENTARY NOTES
This contract was not fully funded resulting in its ending prior to scheduled completion of the effort. As a result, the technical documentation produced is determined to be publishable as a Technical Memorandum.

14. ABSTRACT
Many current social network analytic methods work by analyzing a static aggregate graph, which provides a limited view of the structure and behavior of real-world social networks. Social networks in reality are dynamic and evolve over time as people join or leave the networks and new connections form. This work investigates developing dynamic social network analysis (DSNA) methods to explicitly model time and heterogeneity. It focuses on two objectives: (1) Dynamic SNA metrics and methods which take time into account; (2) Predictive methods for modeling and predicting how individuals and groups change over time.

15. SUBJECT TERMS
Dynamic network analysis, machine learning, graph mining, time-series analysis, social network analysis

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			**TODD WASKIEWICZ**
U	U	U	UU	20	19b. TELEPONE NUMBER (*Include area code)* **N/A**

Contents

List of Figures ii

List of Tables iii

1. Project Motivation and Overview 1

2. Project Status 2

3. Metrics in Dynamic Networks 2

3.1. Centrality in Dynamic Networks 2

3.2. Time-aware Dynamic Centrality 3

3.3. Dynamics-aware Social Proximity in Networks 4

4. Predictive Models for Dynamic Networks 6

4.1. Understanding Social Dynamics 6

4.1.1. Modeling Retweet Behaviors 6

4.1.2. Extracting and Analyzing Social Dialogues 7

4.2. Dynamic Network Prediction 10

5. Developing benchmark data sets 11

6. Conclusion 12

7. Papers related to this effort 13

References 14

List of Figures

Figure 1: Example network. (a) Snapshots of the network showing only connected nodes at times t1; t2; t3 and t4. (b) A static network that aggregates snapshots into a single network.....3

Figure 2: Comparison of importance metrics on Hep-Th data set with different partitions of data. Table on the right lists the metrics used in the evaluation. 4

Figure 3: Predicting user activity (which URLs users will share) on Digg and Twitter. Results are reported as %lift over baseline, which uses (unweighted) friends activity to predict user activity.....5

Figure 4: How to extract dialogues from Twitter streams.....8

Figure 5: Social network. We clearly see strongly connected communities which are loosely connected. The colors represent communities found by standard modularity clustering techniques. 9

Figure 6: Example sub-communities. These structures look more what a standard social network is expected to look like.....10

List of Tables

Table 1: Ratio of users whose overall retweet behavior was best explained by each of the four models. 7

Table 2: How many models were needed to "best" fit the observed behavior of a user? As we can see, many users required a combination of three or four of our models to best explain their behaviors. 7

Table 3: Break-up of Tweets by categories 8

Table 4: Distribution of different dialogue sizes and the number of tweets in those dialogues. 9

1. Project Motivation and Overview

Many current social network analytic methods work by analyzing a static aggregate graph, which provides a limited view of the structure and behavior of real-world social networks. Social networks in reality are dynamic and evolve over time as people join or leave the networks and new connections form. Analyzing a static aggregate of such a network will at best provide an analyst with a historical view of what happened with the network at the time the data was collected, rather than provide the predictive power on how it may look tomorrow or in the future. At worst, such aggregate networks provide a completely skewed view of the true dynamics of a social network to the point where social network analysis will identify the wrong people as influencers or leaders in the network, thereby wasting the valuable time of analysts and leading to targeting the wrong people for further study. Finally, social networks are complex networks which often contain multiple types of relationships and entities—for example people can be related through going to the same event, staying at the same motel, working in the same group, etc. Current Social Network Analysis (SNA) techniques cannot readily handle these complexities and the collapsing of the complex networks to simpler homogeneous (single-entity, single-relation) networks lose significant and crucial information, again leading to skewed analytic results.

In particular, we will focus on developing new and novel ways to analyze dynamic networks to address the shortcomings of current SNA techniques on static aggregate networks including:

1) Who influences whom? Static aggregate networks provide skewed answers.
2) Enhanced community detection algorithms to find communities in dynamic networks (a hopeless task on the static aggregate network).
3) Develop predictive models to track and forecast the evolution of communities and individuals: how does influence and communities change over time.

The research proposed in this effort seeks to directly address these three short-comings by researching and developing dynamic social network analysis (DSNA) methods to explicitly model *time* and *heterogeneity*. We will specifically be focusing on three objectives:

1) Develop dynamic SNA metrics and methods which take time into account.
2) Develop predictive methods for modeling and predicting how individuals and groups change over time, both leveraging the new metrics developed as well as using current SNA metrics as applicable.
3) Extend the above two techniques to be directly applicable on heterogeneous networks.

The result of this research will be threefold: first, our dynamic social network metrics will take time into account and provide improved identification of influencers and leaders in a social network; second, we will develop new predictive models which provide estimates of how individuals and groups will evolve over time, giving analysts crucial situational awareness and the ability to be pre-emptive rather than reactionary. Our third objective will further improve the efficacy of out methods by being able to directly analyze the complex networks rather than rely on initial simplifications to homogeneous networks. The third objective fits directly into the work path of objectives 1 and 2 and is therefore not a separate task.

More concretely, our planned effort consisted primarily of these two high-level tasks:

1) Develop and evaluate dynamic SNA metrics and methods which take time into account. Specifically, we will extend the Bonacich Centrality metric to take time into account, then use the new metric to develop a community detection method.
2) Develop and evaluate predictive methods for modeling and forecasting how individuals and groups change over time, both leveraging the new metrics developed as well as using current SNA metrics as applicable. We will model this as a machine learning task where we have a stream of metrics over time, where the task is to infer a predictive model which predicts with some accuracy how the metrics will change in the next time step(s).

2. Project Status

The effort started in March 2011 and we had the EVOLVE kick-off meeting at AFRL on May 13, 2011. We have been making good progress since the beginning of the project and the work produced follows the original timeline. As such, we are on track with respect to what we had planned at this stage in the effort.

The effort ended early due to the acquisition of the prime contractor.

The rest of the document outlines the work performed for each of the two primary tasks.

3. Metrics in Dynamic Networks

3.1. Centrality in Dynamic Networks

Real-world social networks are dynamic in nature, since their topology can change over time with addition or removal of edges. Figure 1 shows four snapshots of a hypothetical dynamic network, with only connected nodes displayed. A common method to analyze such a dynamic network is to create a static network, shown in Figure 1(b), that aggregates edges observed at all times. However, aggregation loses important temporal information that can help elucidate the dynamic structure of the

network. The toy example could represent travelers who interact with others in different parts of the world. If one person becomes infected (with a pathogen, or receives some information), he can transmit the pathogen (or information) to temporal neighbors in other places. Whether or not the disease (information) will spread, how far and how quickly, who should be immunized to stop it, depends on **both in the nature of interactions between people in a social network and on how the edges in the network evolve over time**. Treating the network as a static aggregate of all edges leads to wrong answers to these questions.

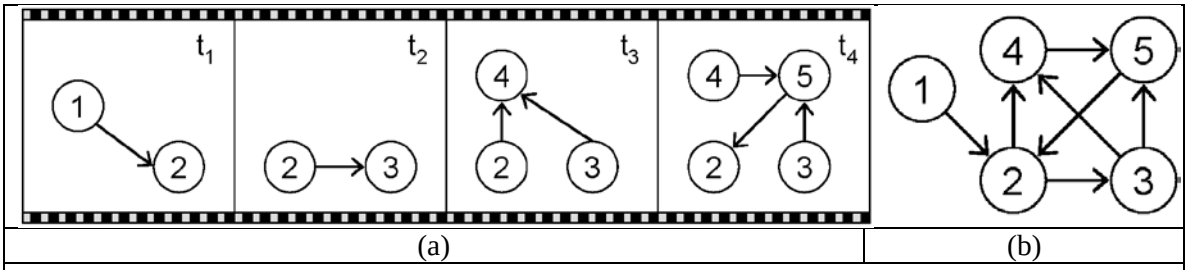


Figure 1: Example network. (a) Snapshots of the network showing only connected nodes at times t1; t2; t3 and t4. (b) A static network that aggregates snapshots into a single network.

Dynamic topology will affect how information flows on a network through interpersonal interactions. For a flow to reach one node from another in a dynamic network, there must exist a path that connects the source and destination nodes through intermediaries at different points in time. Consider a walk from node 1 to 5 in Figure 1. In the static network, there are three acyclic paths from 1 to 5: 1→2→4→5, 1→2→3→4→5, and 1→2→3→5. Not all these paths are physically realizable, however. A walk cannot go from node 1 to node 2 at t1 to node 4 at t2, because an edge 2→4 does not exist at t2. There exists only one path a walk can follow over the period t1–t4, namely 1→2→3→4→5.

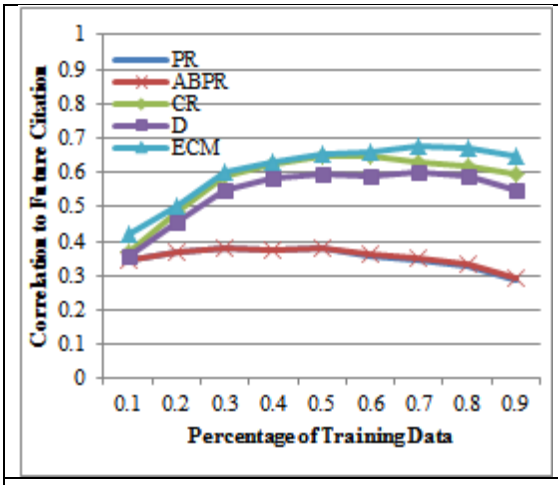
Using this intuition we recently introduced a novel generalization of Bonacich's Alpha-centrality for dynamic networks. It measures centrality of a node by the number of paths that connect it to other nodes through time-dependent edges. A distinctive feature of this metric is that it is parameterized by factors that set both time and length scale of interactions. These parameters can be estimated from data in some cases. In an independent evaluation performed this year at MIT Lincoln Labs, Dynamic Centrality was shown to outperform other metrics in dynamic network analysis.

3.2. Time-aware Dynamic Centrality

Like many other metrics, Dynamic Centrality suffers from **recency bias,** failing to recognize important new nodes that have not had as much time to accumulate links as their older counterparts or temporal order in which links are created.

We studied the problem of time-aware ranking in dynamic networks, specifically citation networks in which nodes are scientific papers and edges are citation links to older papers.

We proposed a time-aware version of *dynamic centrality* that properly discounts older papers while still taking the dynamic nature of the network into account.



Metric	Remark
D	degree = ID + OD
ID	in-degree
OD	out-degree
PR	PageRank [1]
ABPR	Age-based PageRank [2]
CR	CiteRank [5]
ECM	Effective Contagion Matrix (Sec. II-B)
RAM	Recency Adjacency Matrix (Sec. II-A)
SVRECM	ν-SVR [15] w/ D,ID,OD,PR,ABPR,ECM
SVRRAM	ν-SVR [15] w/ D,ID,OD,PR,ABPR,RAM

Figure 2: Comparison of importance metrics on Hep-Th data set with different partitions of data. Table on the right lists the metrics used in the evaluation.

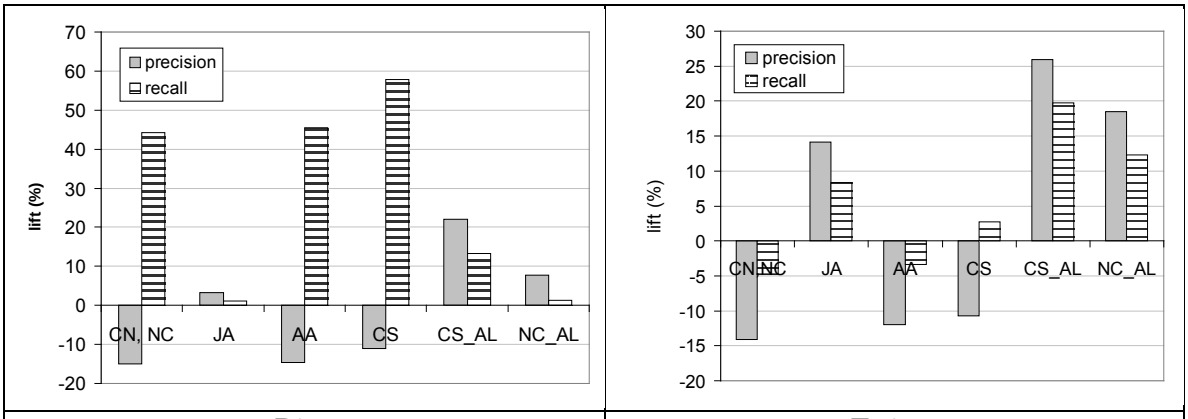
We evaluated our approach on large real-world scientific papers citation networks by seeing how well it predicts papers' future importance. Figure 2 shows prediction results for the HEP-TH data set (10 years of papers from the theoretical physics section of arxiv.org). Time-aware dynamic centrality metric (ECM in Figure 2) is more appropriate for identifying important papers that will attract more citations in the future.

3.3. Dynamics-aware Social Proximity in Networks

The structure of social networks contains information useful for predicting human activity. People who are "close" in some sense in a social network are more likely to perform similar actions than more distant people. In a recent work we used social proximity to capture the degree to which people are "close" to each other within a social network. In addition to standard proximity metrics defined for the link prediction task, such as neighborhood overlap (i.e., number of common neighbors), we introduced novel social proximity metrics that take into account the nature of interactions.

One can think of social proximity as measuring how readily information can be exchanged between two people even in the absence of a direct connection between them. The greater the number of paths connecting the two people, the greater the potential for information exchange; therefore, the closer they are. However, the degree to which information can reach one person from another depends not only on network topology, but also on how information (or influence) *flows* through the network. Consider a network in which people communicate via phone calls. Each person chooses one of her friends and places a call to her. Such one-to-one interactions can be modeled as a random walk; therefore, metrics based on the random walk, such as conductance, are appropriate as a measure of social proximity. However, the spread of information in social media is fundamentally different and cannot be modeled as a random walk. Rather than picking

one neighbor to transmit a message to, a user *broadcasts* the message to all her neighbors. Other examples of such one-to-many processes include epidemics and innovation spread. However, a social media user's capacity to respond to an incoming message is limited by her finite attention, which she must divide over all her friends. As a consequence, the more friends a user has, the less likely she is to respond to an arbitrary message from a friend. This alters the character of the flow and, therefore, how close two people can be considered to be. We have been able to quantify this effect of divided attention on the information spreading behavior on Twitter.



Digg | Twitter

Figure 3: Predicting user activity (which URLs users will share) on Digg and Twitter. Results are reported as %lift over baseline, which uses (unweighted) friends activity to predict user activity.

We recently proposed novel metrics for social proximity that take attention-limited nature of information flow in social media into account. We evaluate the metrics on the activity prediction task, specifically to predict which URLs users will share on Digg and Twitter. These social media sites allow users to post URLs to online content, and other users to share them with others by voting for them (on Digg) or retweeting them (on Twitter). Both sites also allow users to follow activities of friends, creating a directed network we refer to as the follower graph. Friends' activity has been shown to be a useful predictor of user activity in social media. People tend to vote for stories their friends vote for on Digg, use the same tags and favorite the same images as friends on Flickr, and so on. We claim that social proximity can help better predict user activity. Users who are close to each other in the follower graph are more likely to act in similar ways because they share the same information, have similar tastes and attributes, or participate in the same community. Moreover, knowing the actions of some people allows us to better predict the actions of others who are close to them in the network. In a series of experiments using public Digg and Twitter data, we demonstrated that taking into account social proximity leads to better predictions, but only for attention-limited measures of proximity. These findings suggest an important role that attention plays in social media interactions.

4. Predictive Models for Dynamic Networks

This part of the effort is focused on studying how dynamic network analysis metrics change in time and leveraging this information to predict network trends. Specifically, we will study how temporal information can be used to improve our ability to predict the evolution of influential nodes and groups. We will cast this problem as a machine learning problem, where we induce a predictive model which takes as input the last *k* observations (in time) and forecasts the most likely next value for a specific metric value such as the influence of a specific individual in the network.

4.1. Understanding Social Dynamics

4.1.1. Modeling Retweet Behaviors

We studied what drives certain information diffusion processes in social media [Macskassy 2011]. In particular, we studied a set of Twitter users over a period of a month and sought to explain the individual information diffusion behaviors, as represented by retweets, in this domain.

We hypothesized that knowing more about the user and the content would allow us to develop richer models which would take profiles and tagging into account. Specifically, we took an approach to tag Tweets with Wikipedia categories and aggregate these tags for a particular user to generate a topics-of-interest profile for users [Michelson and Macskassy, 2010]. We used these profiles to model retweeting behaviors based on similarities between users and the tweets they retweeted.

We explored four retweeting models, two of which were based on user profiles. The models for predicting retweeting were:

1) **General** model based only on time:
$$P_{\text{gm}}(x) = 0.2 \cdot time(x)^{-1.15}$$
2) **Networking** model, where the only factor was whether users had communicated within the last 24 hours:
$$P_{\text{net}}(x) = P_{\text{gm}} \cdot \left[\alpha \cdot P(x \mid recent(x)) + (1-\alpha) * P(x \mid !recent(x))\right]$$
3) **Topic** model, where the similarity of a user's profile and the topic of a tweet was the key indicator:
$$P_{\text{topic}}(x) = P_{\text{gm}} \cdot P_{\text{ts}}(x \mid \text{sim}_{\text{T}}(x,u))$$
4) **Profile** (or homophily) model, where the similarity between a user's profile and that of the profile of the Twitter user originating a tweet was the key indicator:
$$P_{\text{profile}}(x) = P_{\text{gm}} \cdot P_{\text{ps}}(x \mid \text{sim}_{\text{P}}(x,u))$$

We found that indeed the homophily-based propagation models were better at explaining the majority retweet behaviors we saw in our data as shown in Table 1.

General	Network	Topic	Profile
4%	27%	25%	44%

Table 1: Ratio of users whose overall retweet behavior was best explained by each of the four models.

When digging deeper, however, we found that all four models were used at different times and that user retweeting behaviors, when considering each retweet, were best explained by multiple models. In fact, we found that we got the best overall fit to the data if user behaviors were a modeled as a combination of all four models as shown in Table 2.

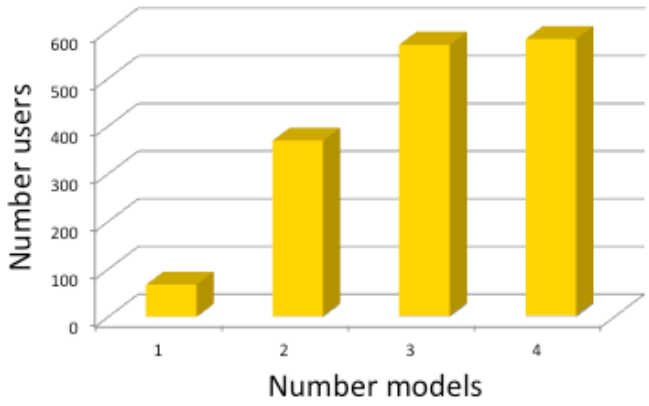


Table 2: How many models were needed to "best" fit the observed behavior of a user? As we can see, many users required a combination of three or four of our models to best explain their behaviors.

This work is a first step into exploring how to leverage content to generate profiles and context in social media, in order to get a deeper understanding of what drives people to propagate or diffuse information. Specifically, we focused on modeling individual microcosm behavior rather than general macro-level processes.

4.1.2. Extracting and Analyzing Social Dialogues

We studied in some detail the behavior and dynamics of dialogues in Twitter [Macskassy 2012]. In particular, we wanted to understand user interaction behaviors, the characteristics of the dialogues people where having and the structure of the emerging social network generated by these interactions.

The first part of out study focused on how to define and extract dialogues. We show in Figure 4 our process for extracting dialogues, where the only variable is *k*, the number of minutes between an observed direct link (a tweet mention) between two users. For example, a dialogue between A and B occurs *if and only if* A mentions B, then B mentions A and *then A mentions B again*, where the maximum delay between mentions is *k* minutes. We explored different values of *k* (from 1 to 9 minutes) and saw qualitatively the same results (fewer dialogues, but same patterns), so we used *k=5* in our study.

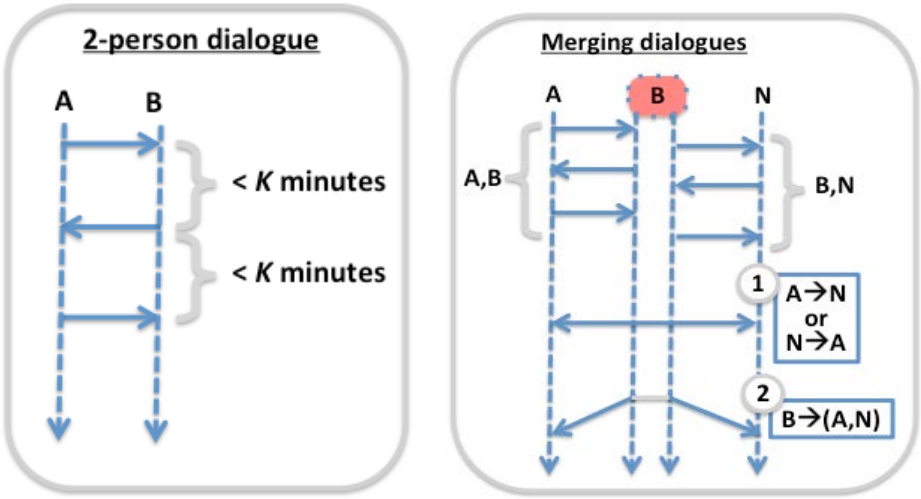


Figure 4: How to extract dialogues from Twitter streams

We found that most people either do not have dialogues or spend about 10% of their Twitter activity in direct interaction with other users. However, we also found that 13% of all tweets in our data were dialogues as shown in Table 3. Of these dialogues, we found that 12% were converted from mentions.

Tweet Category	Number Tweets	Overall Ratio
Dialogue	66,812	0.13
Retweet	93,319	0.19
Mention	154,177	0.31
Tweet	183,748	0.37
Total	498,056	1.00
Conversion	20,155	0.12

Table 3: Break-up of Tweets by categories

We found that the vast majority (over 92%) of dialogues were between two people; about 6% of dialogues were between three people with marginal fractions for larger groups (see Table 4).

Size	Number	Ratio	Avg. Num. Tweets
2	18, 619	92.37%	4.9
3	1, 232	6.11%	8.5
4	181	0.90%	12.7
5	83	0.41%	19.4
6	27	0.13%	36.5
> 6	13	0.07%	> 60

Table 4: Distribution of different dialogue sizes and the number of tweets in those dialogues.

When analyzing the dialogues, we saw a very strong trend for dialogues involving larger number of people tended to not be well-connected although reciprocity was always very high. In other words, although the active dialogue included many people, most explicit interactions were along a few direct mentions. However, we also found that users were very equitable in their interactions, giving and receiving in equal amounts. Interestingly, we found that users were in dialogues with many different people over time but still tended to primarily interact with only a few. This suggested that while the social network would be relatively large, there would be clear strong communities of smaller sizes. We see this clearly in Figure 5.

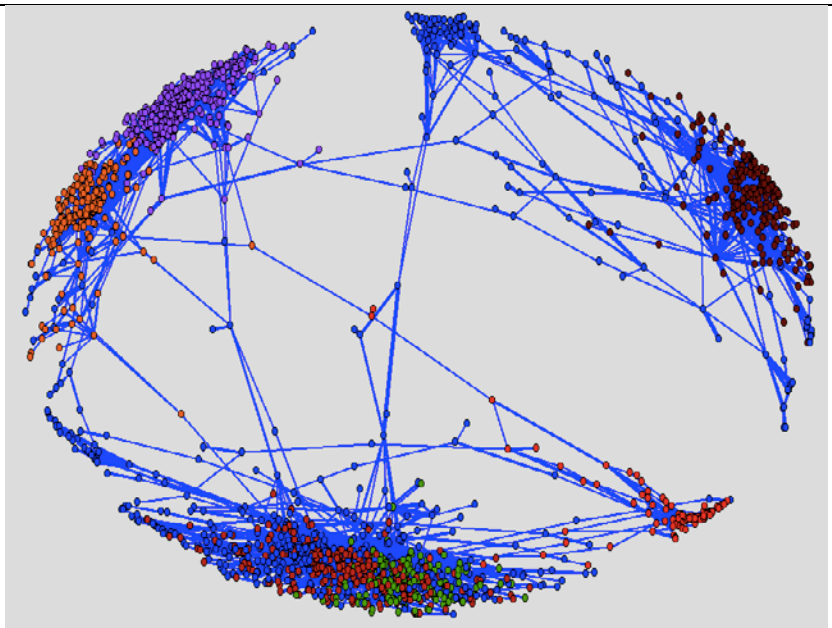


Figure 5: Social network. We clearly see strongly connected communities which are loosely connected. The colors represent communities found by standard modularity clustering techniques.

Taking a closer look at the sub-communities identified, we found that these looked like one would expect from a social network (see Figure 6).

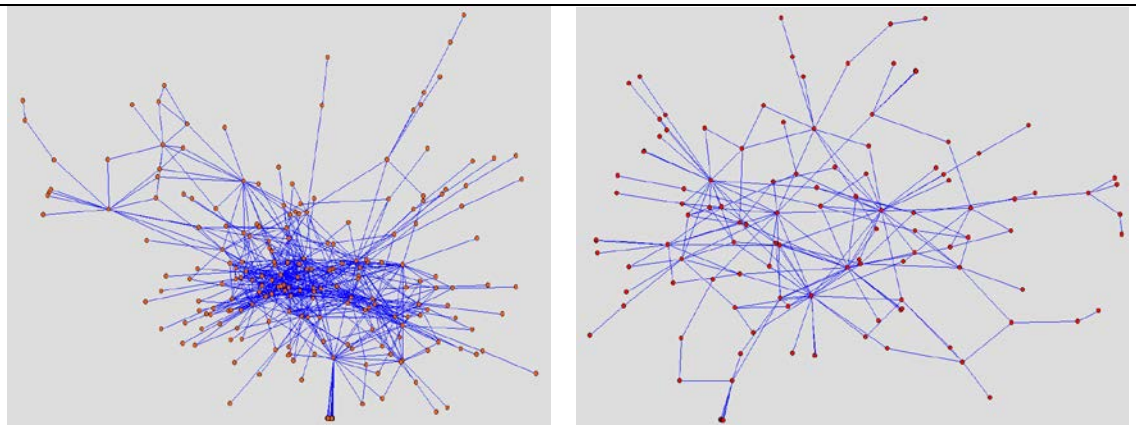


Figure 6: Example sub-communities. These structures look more what a standard social network is expected to look like.

Clearly this work is still indicative of how we should think about and analyze dynamic networks, but it is a first step towards understanding how to extract salient social networks from dynamic interaction streams such as Twitter.

4.2. Dynamic Network Prediction

Per our planned tasks, we started exploring explicit predictive models using machine learning towards the end of Year 1 (and towards the end of the funded effort).

Specifically, we started preliminary work on framing evolution as a machine learning problem, where we want to predict future network behavior. Our first study focused on predicting whether a community was likely to continue, split or disappear in the next time-step (or whether a node was likely to stay or leave a community in the next time-step). In order to do so, we identified two data sets which seemed interesting: the Enron email data set and the World-trade data set. These two data sets were interesting because we have some notion of what the ground truth looks like based on historical facts (world trade: which countries have traded strongly with each other over the past 40 years; Enron: what are the cliques of people speaking with each other over a two year span). The World-trade data comes in yearly snapshots so the time-step is pre-defined. The Enron email data is a stream (each email has a unique time-stamp), so we used one-month snapshots to get a fair amount of snapshots to analyze.

The process we used for our machine learning was as follows:

1) For each snapshot, use community detection (modularity clustering) to identify the communities in that time-step.
2) For each community, compute various node- and community-metrics such as centrality scores. For nodes, we computed metrics such as closeness, betweenness, degree, number of closed triangles, etc. For communities, we computed metrics such as density, number of triangles, in-degree and out-degree (total number of internal edges vs. number of edges going to other communities).
3) For each snapshot, we labeled each the community based on what happened in the subsequent step. For community A in step ***t***, we found community B in step ***t+1*** with the highest overlap.
    a. If the overlap was more than 65% in both A and B, then we labeled this as a **continuing** community.
    b. If more than 65% of A moved into B, then we labeled A as **merging** into B.
    c. If A had two or more sub-groups (each more than 30% of A) continuing, then we said that A **split**.
    d. Otherwise was said that A **dissolved.**
4) We did a similar process with each node, labeling it either as **staying** in a community or **leaving** a community.
5) At each time step, we had labels for each community and each node. We explored different machine learning techniques to understand whether we could do any learning in this setup.
    a. We used the metrics at time step ***t*** to predict next step.
    b. We used the trend over the last 1, 2 and 3 time-steps (for each metric) as well. This is a standard *sliding window* approach in machine learning.

Our results, though preliminary, indicate that we can do some learning in this setup. Specifically, we were able to get AUC scores in the 0.6 to 0.9 range in these prediction tasks. However, we do not yet feel comfortable reporting hard results as these numbers are still early and we need to better understand what they mean and how reliable they are. We intend to strengthen this work and generate a technical paper within the next two months, with the aim of submitting it to a technical venue.

5. Developing benchmark data sets

Finally, one of the core risks identified in our proposal was the dynamic network analysis as we perform it in this effort has a distinct lack of data sets. We spent time identifying potential data sets, including ways of taking existing data sets and converting them into dynamic network benchmark data sets.

We looked at a variety of data sets, some of which we used in our work, and some of which is still work-in-progress.

Twitter data: We have been collecting a variety of Twitter data for different needs. For example, we have looked at (1) information diffusion, where the goal was to

understand how URLs diffuse in the network; (2) psychographic profiling of aggregate streams, where we aggregate tweets and generate a set of concepts that those tweets discuss—this was used to understand retweet behaviors; (3) social dialogues and how they induce social networks—this case is probably closer to the spirit of dynamic networks we wanted to explore in this effort. While we continue with these data and now better understand how to extract networked data, it cannot be used as a public benchmark data set due to Twitter terms of use.

Financial data: We have identified a site: www.insider-monitor.com, which contains information about insider trading. While we downloaded data from this site, we found the networked data to be too sparse to be really useful by itself.

Email data: The Enron email data set is small (150 users), but very rich and covers a 2 year period. We have explored this data in some detail and it turns out to be a very good small-scale network to develop new methods for.

World-trade data: We acquired data on global trade between countries. This data is a yearly snapshot since 1962 of how much trade (import and export) was done between different countries. In addition, these are split up by industry codes. We have started exploring the usefulness of this data at a global scale and hope to have results within the next couple of months. As with the Enron data, the number of nodes is in the low 200's and so the network is relatively small.

Movies: This data set comes from IMDb and is rich in the sense that it has a long history, it has different types of nodes, relations and attributes. We have not yet looked at this for analytics or predictions. It is not immediately clear whether this is an interesting data set.

SNAP: The Stanford Network Analysis project have numerous data sets, some of which are dynamic in nature. The dynamic networks are citation networks as well as some news (memes and twitter). While we have looked at citation networks already, it is not immediately clear that the other networks are good benchmark data for this effort.

In summary, we have identified numerous potential benchmark data sets. Part of making these benchmarks also include how they are used and the results we obtain from them. This task is therefore work-in-progress, which we expect to continue in a future research effort.

6. Conclusion

The overarching goal of this effort is to improve our understanding and technologies for analyzing and managing dynamic networks. To this end, we have focused on three general tasks: (1) Developing new metrics in social networks to identify central people in dynamic networks. These metrics can also be used in other applications such as community detection and diffusion; (2) we have looked at developing richer analytic and predictive models of social behaviors in dynamic social networks such as predicting whether a person is about to leave a community

or retweet a particular post; and (3) creating benchmark data sets so that we have a principled way of analyzing and comparing our work to monitor progress.

We have made good progress in all three general tasks, showing how our dynamic metrics provide insight and ranking of people in ways which are lost when not taking time into account. Our metrics use not only time, but is also informed by social theories such as attention (eg., Dunbar's number) and we have shown how attention and similarity are strong indicators of similar activities.

Our work on predictive modeling was only starting, but we already made good progress on understanding some of the fundamentals such as representation, leveraging node attributes and we were able to get initial preliminary results on predicting community and node behaviors. We started analyzing social networks to understand what makes people retweet and we showed how homopholy was a strong indicator. We further developed initial predictive models of whether communities were about to split or whether nodes were about to leave, using past metrics.

Finally, we spent some time looking for and creating benchmark data sets which we could use to monitor progress. For example, although Twitter has a lot of public data, it is not immediately clear how that can be turned into a salient dynamic network. We looked at extracting social dialogues as one method. We also looked in other domains such as email (the public Enron data set), movies (the freely available IMDb data), bibliometrics (many data), news (how words co-occur), world trade (publicly available) and finances (SEC trading, insider trading data). At the end we focused on smaller and richer data (Enron, World trade) as well as the large networks we could generate from Twitter. These were chosen as they align well with our larger research agenda in social media.

7. Papers related to this effort

Ghosh, R.; Kuo, T.-T.; Hsu, C.-N.; Lin, S.-D.; and Lerman, K. 2011. Time-aware Ranking in Dynamic Citation Networks. In *COMMPER 2011: Mining Communities and People Recommendations, Data Mining Workshops at ICDM*, December.

Hodas, N. and Lerman, K. 2012. How Visibility and Divided Attention Constrain Social Contagion. Submitted to *Social Computing Conference*.

Lerman, K.; Intagorn, S.; Kang, J.-H.; and Ghosh, R. 2012. Using Social Proximity to Predict Activity in Social Networks. Submitted to *ECML/PKDD*. Also, presented as poster at the Int. Conf. on World Wide Web.

Macskassy, S. A. 2011. Why do people retweet? Anti-homophily wins the day! In the Fifth International Conference on Weblogs and Social Media (ICWSM).

Macskassy, S. A. 2012. On the Study of Social Interactions in Twitter. To appear in the Sixth International Conference on Weblogs and Social Media (ICWSM).

Steeg, G. V.; Ghosh, R.; and Lerman, K. 2011. What stops social epidemics? In Proceedings of 5th International Conference on Weblogs and Social Media.

References

Lerman, K.; Ghosh, R.; and Kang, J.-H. 2010. Centrality Metric for Dynamic Network Analysis. In *Proceedings of KDD workshop on Mining and Learning with Graphs (MLG)*, July.

Michelson, M., and Macskassy, S. A. 2010. Discovering users' topics of interest on twitter: A first look. In Proceedings of the Workshop on Analytics for Noisy, Unstructured Text Data (AND).