DHS Summer Internship Report:

Leveraging Temporal Features for Link Prediction in Communication Networks

Tom Yu Ouyang
Massachusetts Institute of Technology
Cambridge, MA 02139
ouyang@csail.mit.edu

September 4, 2007

1 Introduction

The work described in this report was performed by Tom Yu Ouyang during his summer 2007 internship at the Lawrence Livermore National Laboratory (LLNL). Tom was a summer student at the Institute for Scientific Computing Research (ISCR) at LLNL, as a Department of Homeland Security (DHS) Graduate Research Fellow, working under the supervision of Dr. Tina Eliassi-Rad. The internship lasted 10 weeks, and satisfied his summer internship requirement for the 2006 DHS Fellowship.

This report will address the following four topics: 1) the internship project and the student's role in the work, 2) the achievements made during the internship, 3) how the internship experience impacted the student's academic and career planning, and 4) areas of research that would help DHS accomplish its mission and goals.

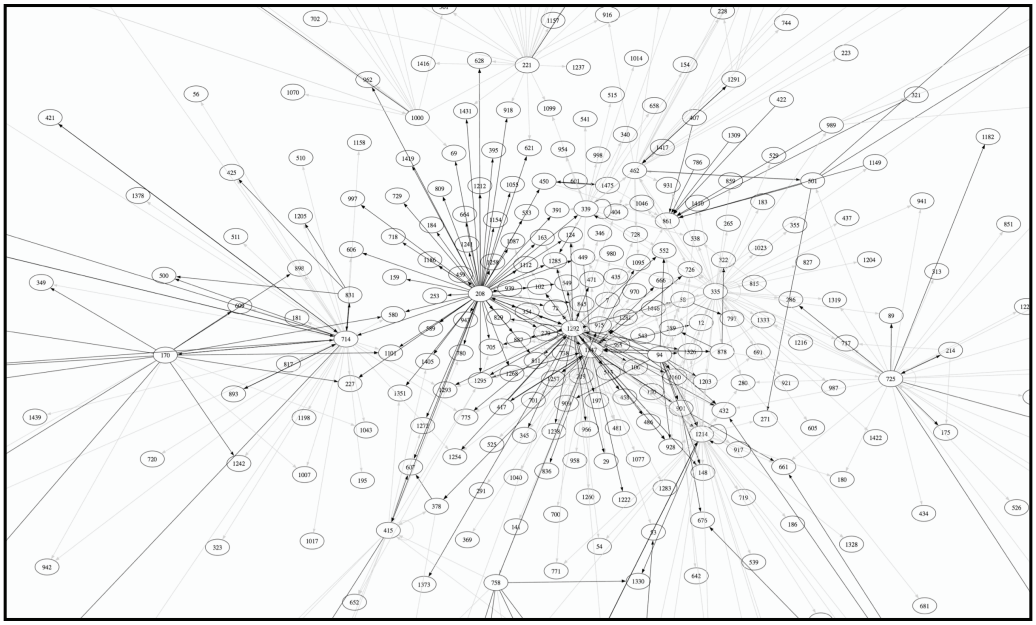

Figure 1: The Enron email communication graph. Nodes and links represent individuals and email communications, respectively. Darker edges indicate more recent links.

2 Link Prediction in Communication Networks

An enormous amount of communication data is accumulated every day in the form of email messages (e.g., see Figure 1), telephone communications, and online social interactions. The ability to monitor and predict future interactions in such relational data (often represented as networks) can be an essential tool in improving national security. For example, it may be useful for law enforcement to determine with whom a suspected terrorist will communicate in the near future, or to model the flow of information within an organization in order to detect fraud or other criminal activities. Employing human experts to analyze these data sets can be prohibitively expensive due to their vast size and their time-evolving nature. Instead, we need to develop intelligent tools that can automatically analyze these types of dynamic networks.

There has been a great deal of research on pattern mining and prediction in social networks [7, 8, 10, 11]. In this project, we focused on the problem of link prediction. In particular, predicting with whom an individual will communicate in the near future based on recently observed activity. Most current approaches to this problem utilize graph-theoretic measures, aggregating links into

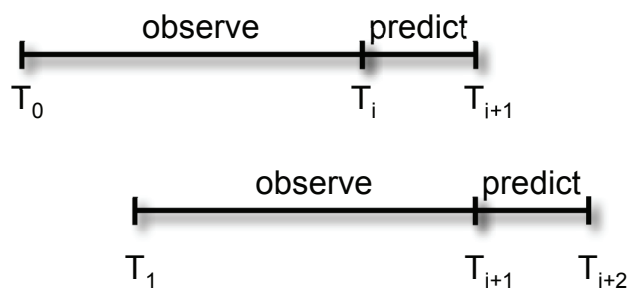

Figure 2: Time frames for observations and predictions. The system predicts the communication pattern in the PREDICT period based on the observed communications in the OBSERVE period.

a static network even though the underlying structure may vary over time [3, 13, 6]. As a result, they largely ignore temporal properties of the graph which may be relevant to link formation. Our hypothesis is that we can improve link predictions in a dynamic network by modeling the temporal aspect of the network.

2.1 Features for Link Prediction

We use a supervised statistical learning approach to predict with whom an individual will communicate in the PREDICT time period based on the recent network activity observed during the OBSERVE time period (see Figure 2). Our system takes advantage of traditional graph-theoretic measures as well as temporal features that model the dynamically evolving nature of the communication network.

In our representation of the network, nodes correspond to individuals while directed edges correspond to email messages. The set of features listed below are generated for each pair of nodes in the communication network. For our purposes, i represents source or sender and j and the target or recipient. $\Gamma(k)$ denotes the set of neighbors of node k (regardless of the direction of links to the neighbors).

2.1.1 Graph-Theoretic Features

These features were derived based on a static representation of the communication network:

- Jaccard Coefficient: The number of shared neighbors of two nodes, normalized by the total number of neighbors of the two nodes.

$$Jaccard := \frac{|\Gamma(i)| \bigcap |\Gamma(j)|}{|\Gamma(i)| \bigcup |\Gamma(j)|} \tag{1}$$

- Minimum Common Neighbor Degree (MCND): The minimum degree of the shared neighbors of two nodes. This feature is similar to the Adamic/Adar coefficient [1], and captures the tendency for a common neighbor with a smaller number of total neighbors to be more significant.

$$MCND := \min_{k \in \Gamma(i) \bigcup \Gamma(j)} (\Gamma(k)) \tag{2}$$

- Katz Coefficient: A weighted sum of the number of directed paths between two nodes. In our system, we use an *unrolled* version of the Katz measure that encodes a separate feature for each path length, l. We use a perceptron learning algorithm to determine the appropriate weights β^l.

$$Katz := \sum_{l=1}^{N} \beta^l |paths^l_{i,j}| \tag{3}$$

- Node Identity: The role or identity of the pair of nodes i and j (e.g., `Assistant` $\rightarrow$ `Manager`).

2.1.2 Temporal Features

These features attempt to capture the time varying nature of the network. In our model, we discretize elapsed time into a set of n partitions $\{p_1, p_2, ..., p_n\}$. For example, a set of partitions could be 0-1 days ago, 1-2 days ago, 2-5 days ago, and so on.

- Most Recent Link Direction: The direction (send or receive) of the most recent link between the source i and target j.

- Most Recent Link Time: The elapsed time partition (e.g., p_t) of the most recent link between

the source i and target j.

- Link Pattern: An array of the elapsed time partitions in which there was a link between the source i and target j (e.g., [p_{t1}, p_{t2}, ...]). This is intended to capture any frequent patterns of communications between two nodes over time.

- Forwarding Pattern: The set of identities (e.g., Assistant, Manager) of all nodes that have sent a message to the source node i, conditioned on the identities of the source and target nodes. This features is used to model common forwarding patterns based on node identities.

- Co-recipients: This feature is activated when the source and target nodes i and j were co-recipients of an message with multiple recipients.

- Temporal Graph Features: Temporal variations of the graph-theoretic measures described in the previous section (e.g., Katz) derived from the structure of the communication network at each discretized time partition p_t.

2.2 Supervised Learning Model

We used a discriminative machine learning approach to determine the importance of each feature from the data. In particular, we used a perceptron online learning algorithm [12] that can update its parameters as new data instances are observed without needing lengthy "retraining." This allows the system to rapidly incorporate new information as the graph evolves.

2.3 Evaluation

In order to train and evaluate our model, we used the Enron email dataset [2] that is publicly available online. This corpus contains data from about 150 employees at Enron and a total of about 500,000 messages that was originally made public during the investigation by the Federal Energy Regulatory Commission. For our task, we augmented the email addresses contained in the data set (both TO and FROM the Enron employees) with the label identities (e.g., Manager,

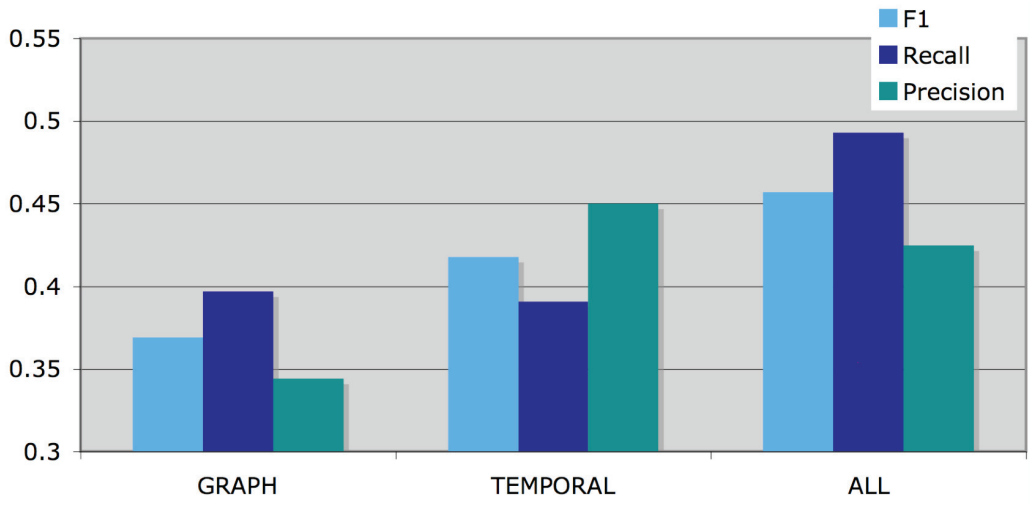

Figure 3: F1 measure, Recall, and Precision for GRAPH (Baseline), TEMPORAL, and ALL feature groups.

Assistant, etc.) that was provided by researchers at UC Berkeley [9]. Since identity labels were not available for all of the individuals in the data set, we used the subset containing of about 1,400 email addresses.

For our experiments, we trained and tested our system on 30,000 emails over the period of 3 months. We used an *observation window* of 1 month and a *predict window* of 1 day, and ignored the content of the messages (i.e. their text), which left us with only the identity of the sender and receiver. Our system learned a model that predicts with whom a given individual would communicate during the PREDICT frame, conditioned on the fact that the individual did communicate with someone in that frame.

We used the first two months of the data to train the link prediction model. The system was then evaluated on its ability to predict the communications that occurred in the remaining month. Figure 3 reports the F1-measure for this task, given by the geometric mean of the precision and recall. The baseline GRAPH uses only graph-theoretic measures, TEMPORAL uses only temporal features, and ALL uses both graph-theoretic and temporal features.

2.4 Discussion

Overall, our results suggest that there is a need for further improvement before our model is ready to be used in real-world applications. However, we observed gains when modeling temporal information in the network. In addition, we found that our performance on predicting future messages between employees who have communicated recently is significantly better than on novel links (i.e., those between two individuals who have not communicated in the last month). Examination of the data does not reveal any prevailing pattern for these anomalous links, which raises doubts as to whether they can be anticipated without more information.

While there has been a great deal of related literature on inference and analysis in social networks, relatively little work has been done on link prediction in dynamically evolving networks at the level of time granularity (i.e., one day) presented in this report. One exception is the work by Lahiri and Berger-Wolf [5] which looks at the problem of structure prediction using frequent subgraphs. They present encouraging results of up to 85% accuracy on data sets including the Enron email network. However, their evaluation differs from ours in their use of a slack variable, which increases the time interval in which a prediction is considered accurate. Furthermore, they also limit their model to frequent patterns and, as a result, their predictions cover only subset of the actual interactions in the network.

2.5 Student's Role

The work described above was performed as part of a two-student team consisting of myself (Tom Yu Ouyang) and Aria Haghighi, under the supervision of Tine Eliassi-Rad. I was involved in all parts of the project including the planning, design, and implementation of the system. In particular, I focused more on the feature selection, training methodology, and performance evaluation aspects.

3 Internship Achievements

- We evaluated a range of features for link prediction in a communication network. In particular, we experimented with two sets of features: graph-based and time-based. The former was based on static representations of the communication network – measured by traditional graph-theoretic metrics. The latter captured the time varying nature of the network.

- We developed a link prediction model that utilizes both temporal and graph-theoretic features. We trained and evaluated our model using the Enron email data set. Our results confirmed our hypothesis: models that combine temporal and graph-theoretic feature sets achieve a higher classification performance than models that use only one of the two feature sets.

- We presented our work to other students and researchers at the lab-wide Summer Student Research Symposium held at LLNL on August 9, 2007.

4 Internship Experience

One of the most memorable aspects about my summer internship at LLNL was the opportunity to meet and get to know the other students and researchers at LLNL. It was truly exciting to meet people from such a wide range of academic and research backgrounds. The student research symposium, in particular, was a great way to learn about the research done by other DHS Scholars and Fellows. It was also an excellent opportunity to share my own work with other LLNL employees.

During my visit, I also had the opportunity to attend some of the frequent seminars and talks on topics ranging from chem-bio detection to information support and management systems for intelligence analysts. I really appreciated the chance to learn about the breadth of work that was happening at LLNL, especially those that were outside of my research area. It was also a wonderful experience working with the Predictive Knowledge Systems group at LLNL. Through weekly meetings with Tina Eliassi-Rad and the other members of the group, I learned a great deal about

their work and the related research in machine learning and knowledge discovery.

The opportunity to work on this project over the past summer has also raised my awareness of national security issues and how they relate to my research and career goals. The internship experience has encouraged me to seriously consider returning for another DHS internship next summer. Furthermore, through learning about the excellent opportunities and stimulating work environment at LLNL, I am beginning to think about the possibility of pursuing a postdoctoral fellowship or research position at a national laboratory after I graduate.

5 Homeland Security Missions and Goals

This project contributes to the DHS mission in several ways. First, our results suggest ways for analysts to automatically predict future interactions between agents in communication networks based on their past behavior. Second, our statistical model of link prediction can be adapted to the task of anomalous link discovery, where we wish to detect links in a network that are statistically unlikely and therefore worth further investigation. Finally, our analysis of temporal features in communication networks suggest ways to improve the accuracy of other graph-based tasks, such as automatic group discovery in dynamically evolving networks [4].

6 Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-ENG-48 (UCRL-TR-234210) and under the U.S. Department of Homeland Security (DHS) Scholarship and Fellowship Program, a program administered by the Oak Ridge Institute for Science and Education (ORISE) for DHS through an interagency agreement with the DOE. ORISE is managed by Oak Ridge Associated Universities under DOE contract number DE-AC05-06OR23100.

References

[1] L.A. Adamic and E. Adar. Friends and neighbors on the Web. *Social Networks*, 25(3):211–230, 2003.

[2] Enron Email Dataset. http://www.cs.cmu.edu/ enron/.

[3] Lise Getoor and Christopher P. Diehl. Link mining: A survey. *ACM SIGKDD Explorations Newsletter*, 7(2):3–12, 2005.

[4] P.S. Koutsourelakis. Unsupervised Group Discovery and Link Prediction in Relational Datasets: A nonparametric Bayesian approach. Technical report, UCRL-TR-230743, Lawrence Livermore National Laboratory, Livermore, CA, 2007.

[5] M. Lahiri and T.Y. Berger-Wolf. Structure prediction in temporal networks using frequent subgraphs. In *IEEE Symposium on Computational Intelligence and Data Mining (CIDM)*, pages 35–42, 2007.

[6] D. Liben-Nowell and J. Kleinberg. The Link prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

[7] Q. Lu and L. Getoor. Link-based classification. *Proceedings of the 20th International Conference of Machine Learning (ICML)*, pages 496–503, 2003.

[8] Joshua O'Madadhain, Jon Hutchins, and Padhraic Smyth. Prediction and ranking algorithms for event-based network data. *ACM SIGKDD Explorations Newsletter*, 7(2):23–30, 2005.

[9] U.C. Berkeley Enron Email Analysis Project. http://bailando.sims.berkeley.edu/enron_email.html.

[10] M.J. Rattigan and D. Jensen. The case for anomalous link discovery. *ACM SIGKDD Explorations Newsletter*, 7(2):41–47, 2005.

[11] D. Roberts and T. Eliassi-Rad. A position paper: Value of information for evidence detection. In *AAAI Fall Symposium on Capturing and Using Patterns for Evidence Detection*, 2006.

[12] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[13] B. Taskar, M.F. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. *Advances in Neural Information Processing Systems*, 16, 2004.