Using Proximity to Predict Activity in Social Networks

Kristina Lerman
Information Sciences Institute
University of Southern California
Marina del Rey, CA 90292
lerman@isi.edu

Suradej Intagorn, Jeon-Hyung Kang, Rumi Ghosh
University of Southern California
Los Angeles, CA
{intagorn,jeonhyuk,rumig}@usc.edu

ABSTRACT

The structure of a social network contains information useful for predicting its evolution. We show that structural information also helps predict activity. People who are "close" in some sense in a social network are more likely to perform similar actions than more distant people. We use network proximity to capture the degree to which people are "close" to each other. In addition to standard proximity metrics used in the link prediction task, such as neighborhood overlap, we introduce new metrics that model different types of interactions that take place between people. We study this claim empirically using data about URL forwarding activity on the social media sites Digg and Twitter. We show that structural proximity of two users in the follower graph is related to similarity of their activity, i.e., how many URLs they both forward. We also show that given friends' activity, knowing their proximity to the user can help better predict which URLs the user will forward. We compare the performance of different proximity metrics on the activity prediction task and find that metrics that take into account the attention-limited nature of interactions in social media lead to substantially better predictions.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

proximity, social networks, activity prediction, social media

1. INTRODUCTION

The structure of complex networks contains valuable information that can be used to identify missing links and predict which new links between existing nodes are likely to be observed in the near future [3, 2, 4]. Given a pair of unconnected nodes, link prediction algorithm calculates a graph-based proximity score between them. Graph proximity measures how readily information can be exchanged by nodes in a network even in the absence of a direct link between them. However, the degree to which node is reachable depends not only on network topology, but also on the nature of interaction between the nodes [1]. One-to-one interactions such as web surfing or phone conversations, can be described as a random walkbut in social media, rather than pick one neighbor to whom to transmit a message, users *broadcast* it to all neighbors. Also, since users' capacity to respond to incoming messages from network neighbors is limited by their finite attention, this may further change the nature of interactions in social media.

We propose proximity metrics that take into account the one-to-many and attention-limited interactions between nodes. We show that structural proximity can help predict URL forwarding activity in social media. When a user tweets a URL, it is is broadcast to all the user's followers, who may in turn retweet it. We investigate how well follower graph-based proximity metrics predict which URLs the user will retweet. We find that metrics that take into account the one-to-many and attention-limited nature of interactions lead to better predictions.

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

2. INTERACTIONS AND PROXIMITY

Table 1: Some of the proximity metrics used for network analysis, including four proposed in this paper

metric	*definition*
CN	$CN = \frac{1}{2}\left[\lvert\Delta\rvert + \lvert\Delta'\rvert\right]$
JC	$\text{JC}= \frac{1}{2}\left[\frac{\lvert\Gamma_{\text{out}}(u)\cap\Gamma_{\text{in}}(v)\rvert}{\lvert\Gamma_{\text{out}}(u)\cup\Gamma_{\text{in}}(v)\rvert} + \frac{\lvert\Gamma_{\text{out}}(v)\cap\Gamma_{\text{in}}(u)\rvert}{\lvert\Gamma_{\text{out}}(v)\cup\Gamma_{\text{in}}(u)\rvert}\right]$
AA	$AA = \frac{1}{2}\left[\sum_{z\in\Delta}\frac{1}{\log(d(z))} + \sum_{z'\in\Delta'}\frac{1}{\log(d(z'))}\right]$
CS	$CS = \frac{1}{2}\sum_{z\in\Delta}\frac{1}{d_{\text{out}}(u)d_{\text{out}}(z)} + \frac{1}{2}\sum_{z\in\Delta'}\frac{1}{d_{\text{out}}(v)d_{\text{out}}(z)}$
CS_AL	$CS\_AL = \frac{1}{2}\sum_{z\in\Delta}\frac{1}{d_{\text{out}}(u)d_{\text{in}}(z)d_{\text{out}}(z)d_{\text{in}}(v)} + \frac{1}{2}\sum_{z\in\Delta'}\frac{1}{d_{\text{out}}(v)d_{\text{in}}(z)d_{\text{out}}(z)d_{\text{in}}(u)}$
NC	$NC = \frac{1}{2}\left[\lvert\Delta\rvert + \lvert\Delta'\rvert\right]$
NC_AL	$NC\_AL = \frac{1}{2}\sum_{z\in\Delta}\frac{1}{d_{\text{in}}(z)d_{\text{in}}(v)} + \frac{1}{2}\sum_{z\in\Delta'}\frac{1}{d_{\text{in}}(z)d_{\text{in}}(u)}$

Intuitively, network proximity measures the likelihood a message starting at node u will reach another node v, regardless of whether an edge exists between them. The greater the number of paths connecting them, the more likely they are to share information, and the closer they are in the network. Proximity metrics used in previous studies [3, 4] include the number of common neighbors (CN), fraction of common neighbors, or Jaccard (JC) coefficient, and the Adamic-Adar (AA) score, which weighs each common neighbor by the inverse of the log of its degree. Table 1 gives their definition us-

ing directed neighborhoods of u and v: $\Delta = \Gamma_{\text{out}}(u) \cap \Gamma_{\text{in}}(v)$ and $\Delta' = \Gamma_{\text{in}}(u) \cap \Gamma_{\text{out}}(v)$. Here, $\Gamma_{\text{out}}(u)$, represents the set of out-neighbors of node u, which in social media corresponds to the set of followers of u. Similarly, $\Gamma_{\text{in}}(u)$ represents the set of in-neighbors (friends) of u. The out-degree of u is $d_{\text{out}}(u) = |\Gamma_{\text{out}}(u)|$ and in-degree is $d_{\text{in}}(u)$.

The likelihood a message will reach v from u depends, however, not only on the number of paths, but also on the nature of the dynamic process by which messages spread on the network [1]. Different dynamic processes will lead to different notions of proximity, even in the same network. Consider first a random walk, or what we call a *conservative* process. Koren et al. [2] introduced cycle-free effective conductance as a measure of proximity. This is a global metric computes the probability a random walk starting at u will reach v through any path in the graph. In most cases we are interested in *local* measures, that depend only on the neighborhoods of u and v. They are not only easier to compute, but also do not require knowledge of the full graph, e.g., the entire Twitter follower graph. To go from u to v, the random walker first needs to pick an edge that will take it u to a common neighbor z it shares with v (which it will do with probability $1/d_{\text{out}}(u)$), then it has to pick an edge that will take it to v (which it will do with probability $1/d_{\text{out}}(z)$). Symmetrizing, we obtain metric CS in Table 1. This measure is almost identical to the metric shown by Zhou et al. [5] to perform best on the missing link prediction task in the electric power grid, router-level Internet graph, and US air transportation networks, all of which have conservative interactions.

People have finite attention, which limits their capacity to respond to incoming stimuli. Social media users divide their attention among all friends, which limits their ability to respond to a specific friend (for simplicity, we assume that attention is evenly divided among friends). This alters the interactions. Now, in order for a message to get from u to a common neighbor z, it must not only go over the correct out-link from u, but z must also pay attention to the in-link, which it will do with probability $1/d_{in}(z)$. This leads to attention-limited conservative metric CS_AL in Table 1.

Now imagine that messages flow via one-to-many broadcasts. For a message to get from u to v, first u broadcasts it to its neighbors, including z, and then z broadcasts it. Probability of getting the message to v is one; therefore, non-conservative proximity NC simply counts the neighborhood overlap. Finite attention can also play a role in non-conservative interactions. Following the logic above, we derive attention-limited version NC_AL. In undirected graphs, it is identical to conservative metric.

3. ACTIVITY PREDICTION

Social media users tend to be similar to their friends, which means that they tend to vote for URLs their friends vote for on Digg or retweet on Twitter, and so on. While friends' activity can be a useful predictor of user's actions, we claim that knowing the local structure of the follower graph can enhance predictive power. In other words, while social media users tend to act like their friends, they are more likely to act like their closer friends.

We evaluate this claim by predicting URL forwarding on Digg and Twitter. The task can be stated as follows: given the follower graph and the URLs that a user's friends forward (retweet), predict which stories the user retweets. We construct a prediction vector p for a user, whose values represent probability a user's friends retweet the i^{th} URL, weighted by each friend's proximity. To compute precision and recall of prediction, we construct a vector of URLs the user actually retweeted. We compare proximity-based prediction to *baseline* that weighs friends' activity uniformly, without regard to proximity to user. We measure performance as improvement over baseline (lift).

Table 2: Evaluation of predictions by different metrics in the Digg and Twitter data sets. Lift is defined as % change over baseline.

	base	CN, NC	JA	AA	CS	CS_AL	NC_AL
(a) Digg							
precision	0.032	0.027	0.033	0.027	0.028	0.039	0.034
recall	0.172	0.248	0.174	0.250	0.272	0.195	0.174
pr lift %	0	-15.0	3.3	-14.7	-11.1	**22.1**	**7.7**
re lift %	0	44.2	1.1	**45.5**	**57.9**	13.3	1.3
(b) Twitter							
precision	0.105	0.091	0.120	0.093	0.094	0.133	0.125
recall	0.094	0.090	0.102	0.091	0.097	0.113	0.106
pr lift %	0	-14.1	14.1	-12.0	-10.7	**25.9**	**18.5**
re lift %	0	-4.8	8.4	-3.4	2.8	**19.7**	**12.3**

We used the Digg data set,[1] which contains voting records of 139K users on 3.5K stories. The Digg follower graph has 70K nodes and more than 1.7 million edges. Our Twitter data set contains retweeting histories of 4K URLs that have been tweeted by 542K users. The Twitter follower graph has almost 700K nodes and over 36 million edges.

Table 2 compares prediction performance of different proximity metrics. Attention-limited versions of proximity metrics result in the greatest lift both in precision and recall. This is because they account for the nature of communication in social media.

Acknowledgments

This paper is based on work funded by the Air Force Office of Scientific Research under contracts 1295GNA276 and FA9550-10-1-0102, and by the National Science Foundation under grant 0915678.

4. REFERENCES

[1] R. Ghosh, K. Lerman, T. Surachawala, K. Voevodski, and S.-H. Teng. Non-Conservative diffusion and its application to social network analysis. Technical report, University of Southern California, Feb 2011.

[2] Y. Koren, S. C. North, and C. Volinsky. Measuring and extracting proximity graphs in networks. *ACM Trans. Knowl. Discov. Data*, 1(3), Dec. 2007.

[3] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci.*, 58(7):1019–1031, 2007.

[4] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, Dec. 2010.

[5] T. Zhou, L. Lü, and Y.-C. Zhang. Predicting missing links via local information. *The European Physical Journal B - Condensed Matter and Complex Systems*, 71(4):623–630, Oct. 2009.

[1] http://www.isi.edu/∼lerman/downloads/digg2009.html