Using Community Information to Improve the Precision of Link Prediction Methods

Sucheta Soundarajan
Cornell University, Dept. of Computer Science
Ithaca, New York 14853
sucheta@cs.cornell.edu

John Hopcroft
Cornell University, Dept. of Computer Science
Ithaca, New York 14853
jeh@cs.cornell.edu

ABSTRACT

Because network data is often incomplete, researchers consider the link prediction problem, which asks which nonexistent edges in an incomplete network are most likely to exist in the complete network. Classical approaches compute the 'similarity' of two nodes, and conclude that highly similar nodes are most likely to be connected in the complete network. Here, we consider several such similarity-based measures, but supplement the similarity calculations with community information. We show that for many networks, the inclusion of community information improves the accuracy of similarity-based link prediction methods.

Categories and Subject Descriptors

H.3.4 [**Systems and Software**]: Information Networks

Keywords

Social Networks, Link Prediction, Communities

1. INTRODUCTION

Network analysis is becoming an increasingly popular research topic in computer science. Because data is often incomplete (e.g., in genetic networks where links are determined experimentally), researchers wish to predict which missing links are most likely to exist. Algorithms for this problem typically assume that 'similar' nodes are likely to be connected [5]. The question, then, lies in defining 'similarity.' In this paper, we consider several methods that use local information to calculate similarity, and supplement them with community information. Over 10 networks, our enhanced metrics typically outperform the original methods.

2. ENHANCED LINK PREDICTION

We believe that community membership information, identified through metadata, community detection algorithms, or other sources, can provide valuable information for link prediction. Consider a metric in which the similarity of nodes a and b is based on their number of shared neighbors. Suppose that we use this metric to analyze a friendship network, and that shared neighbor c knows a from the same school and b from the same workplace, while shared neighbor d knows both a and b from the same sports team. Intuitively, it seems that d should contribute more heavily to the similarity between a and b, because d knows both a and b from the same context.

With this intuition, we modify several base local similarity metrics for link prediction: Common Neighbors (CN), Resource Allocation (RA), Jaccard Similarity, Leicht-Holme-Newman, and Sorensen Similarity [5]. Of these, CN, RA, or one of their modifications is the top performer on each dataset, so we present only these measures here.

For nodes a and b, let $\Gamma(a)$ be the set of neighbors of a, $\Gamma(a,b) = \Gamma(a) \cap \Gamma(b)$, and $d(a)$ be the degree of a. Then:

Common Neighbors: $CN(a,b) = |\Gamma(a,b)|$.

Resource Allocation: $RA(a,b) = \sum_{c \in \Gamma(a,b)} \frac{1}{d(c)}$.

We generate communities using the Louvain method for greedy modularity optimization (Mod) [2], Infomap (IM) [8], and the Link Communities method (LC) [1]. LC generates communities of edges rather than nodes; we thus interpret its results in two ways: first, as a collection of overlapping communities of nodes (each, a "node-community"), and second, as a partitioning of edges (each, an "edge-community").

We consider a variety of modifications to the original local similarity measures, including awarding extra points to pairs of nodes that share many communities and penalizing pairs of nodes that do not share communities. For each of CN and RA, one modification stood out, and so due to space constraints, we present only those here. Because we consider both node-communities and edge-communities, we present two versions of each of these modifications. CN1 and RA1 use node-communities from Mod, IM, and LC, while CNEdge1 and RAEdge1 use edge-communities from LC.

For each of the metrics described below, $C(a)$ and $C(a,b)$ are, respectively, the set of node-communities containing a and the set of edge-communities containing edge (a,b).

- **Common Neighbors 1 (CN1) / Common Neighbors Edge 1 (CNEdge1):** Begin with $CN(a,b)$, and for every neighbor i shared by a and b, $CN1(a,b)$ adds a point for every community that a, b, and i share. $CNEdge1(a,b)$ adds a point if edges (a,i) and (b,i) are in the same edge-community.

$$CN1(a,b) = CN(a,b) + \sum_{i \in \Gamma(a,b)} |C(i) \cap C(a) \cap C(b)| .$$

$$CNEdge1(a,b) = CN(a,b) + \sum_{i \in \Gamma(a,b)} |C(i,a) \cap C(i,b)| .$$

- **Resource Allocation 1 (RA1) / Resource Allocation Edge 1 (RAEdge1):** $RA1(a,b)$ modifies

Copyright is held by the author/owner(s).
WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

$RA(a,b)$ to only consider neighbors i that are in a community with both a and b, and weights i's contribution by the number of communities it shares with a and b. $RAEdge1(a,b)$ considers only shared neighbors i with edges (a,i) and (b,i) in the same edge-community.

$$RA1(a,b) = \sum_{i \in \Gamma(a,b)} \frac{|C(i) \cap C(a) \cap C(b)|}{d(i)}.$$

$$RAEdge1(a,b) = \sum_{i \in \Gamma(a,b)} \frac{|C(i,a) \cap C(i,b)|}{d(i)}.$$

3. EXPERIMENTS AND RESULTS

We test these metrics on 10 datasets: **Amazon** (a book co-purchasing network from Amazon.com [3]), **Grad** and **Ugrad** (portions of the Facebook network corresponding to graduate and undergraduate students at Rice University [6]), **HS** and **SC** (protein interaction networks for humans and a yeast species [7]), **Email** (an e-mail network from the University Rovira i Virgili [4]), **HEP** and **Rel** (collaboration networks from arxiv.org for two fields of physics [3]), **Wiki** (voting network from Wikipedia elections [3]), and **Word** (an experimentally created associative thesaurus [9]). These networks range in size from 503 nodes and 3256 edges (Grad) to $270,347$ nodes and $741,142$ edges (Amazon).

For each network, we perform experiments using 10-fold cross validation, in which 90% of the links are used as training data and the remaining 10% used for testing. For each round of experiments, we use the training data to generate communities using IM, Mod, and LC (each algorithm constitutes a different experiment). Using our metrics, we identify the n most likely links, where n is 10% of the size of the test data, and determine the fraction of these links present in the test data. Our networks may be incomplete (even when data is not withheld), and so even a perfect link prediction metric may not receive a perfect score, because it may predict links that exist in the full data, but do not exist in the incomplete data. Thus, we are chiefly interested in how scores compare across different metrics, not the absolute scores themselves.

Table 1 contains the results of these experiments. To save space, we present the performance of each base metric (CN and RA) and the performance of CN1 (or CNEdge1) and RA1 (or RAEdge1) for the best performing algorithm (e.g., the row for 'Amazon' indicates that the communities from IM allowed CN1 to achieve a precision of 0.374). 'LCE' indicates that CNEdge1 or RAEdge1, with communities from Link Communities, was the best metric. Table 2 contains the average performance of each metric for each community detection method.

These results show that use of community information typically leads to an improvement in precision, sometimes by a large degree, such as with network SC, which saw a 6-fold improvement from RA to RA1. With one exception, the best performing metric is an enhanced metric. On average, the tested modifications improved upon the base metric regardless of community detection method, and CNEdge1 and RAEdge1 give the greatest improvement overall.

Table 1: Precision for Base and Enhanced Metrics

		CN1/		RA1/
	CN	CNEdge1	RA	RAEdge1
Amazon	0.371	0.374 (IM)	0.351	**0.411 (LC)**
Grad	0.552	0.548 (LCE)	**0.72**	0.715 (LC)
Ugrad	0.576	0.675 (Mod)	0.689	**0.724 (IM)**
HS	0.111	**0.157 (LC)**	0.073	0.126 (LCE)
SC	0.194	0.375 (LC)	0.083	**0.471 (LCE)**
Email	0.351	0.369 (LCE)	0.326	**0.384 (LC)**
HEP	0.699	0.804 (LCE)	0.923	**0.928 (LC)**
Rel	0.968	0.968 (all)	0.99	**0.995 (LCE)**
Wiki	0.177	**0.195 (Mod)**	0.142	0.146 (LC)
Word	0.140	0.149 (Mod)	0.147	**0.149 (Mod)**
Average	0.414	0.444 (LCE)	0.444	**0.494 (LCE)**

Table 2: Avg. Prec. for Base and Enhanced Metrics

	Base	Mod	IM	LC	LCE
CN	0.414	0.435	0.436	0.433	**0.444**
RA	0.444	0.455	0.489	0.455	**0.494**

4. CONCLUSION AND FUTURE WORK

These results show that community information often boosts the performance of base metrics, sometimes by a large amount. We wish to identify which factors lead to the success of a particular metric. Which network features can guide selection of an appropriate metric? We are also interested in fine-tuning our definitions: for instance, in CN1, the similarity between two nodes gets a point for every shared neighbor and every shared community. Can we weight these values differently? We have shown that a simple modification boosts accuracy, but more improvement is likely possible.

5. ACKNOWLEDGEMENTS

This work was funded by Air Force Office of Scientific Research grant AFOSR FA9550-09-1-0675.

6. REFERENCES

[1] Y. Y. Ahn, J. P. Bagrow, and S. Lehmann. Link communities reveal multiscale complexity in networks. *Nature*, 466, 2010.

[2] V. D. Blondel, et al. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10), 2008.

[3] J. Leskovec, et al. SNAP website, http://snap.stanford.edu/data/index.html, 2012.

[4] R. Guimera, L. Danon, A. Diaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Phys. Rev. E*, 68, 2003.

[5] L. Lu and T. Zhou. Link prediction in complex networks: a survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, March 2011.

[6] A. Mislove, B. Viswanath, K. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. *WSDM*, 2010.

[7] D. Park, R. Singh, M. Baym, C. Liao, , and B. Berger. Isobase: A database of functionally related proteins across ppi networks. *Nucleic Acids Research*, 2011.

[8] M. Rosvall and C. T. Bergstrom. Maps of information flow reveal community structure in complex networks. *PNAS*, 105(4):1118–1123, January 2008.

[9] G. Kiss, C. Armstrong, R. Milroy, and J. Piper. An associative thesaurus of english and its computer analysis. In *The Computer and Literary Studies.* Edinburgh: University Press, 1973.