CS224W Course Project Proposal: A survey of Network Alignment

Danqi Chen
Stanford University
danqi@stanford.edu

Botao Hu
Stanford University
botaohu@stanford.edu

Shuo Xie
Stanford University
shuoxie@stanford.edu

October 22, 2012

Abstract

This document is a project proposal for the CS231A open course project. It details our plans for contributing to current research in real-time object tracking. Possible datasets, algorithms, readings and evaluation methods are reviewed.

1 Problem Statement

2 Algorithms

3 Data and Evaluation

We will use the data from three large online social networks in our experiments: Twitter, Flickr and Foursquare. On these social networks, the data of user profiles and friendship connections are all public and accessable by crawlers or APIs.

The first graph is the "following" relationships on the Twitter[1], a microblogging service, which has 500 million users (200 million active).

The second graph is the "contact" relationship on Flickr[2], a photo-sharing service, which has 51 million registered members and 6 billion images on Jan 19, 2012.

The third graph is the "Friends" relationships on Foursquare[3], a location-based social network, which has 22 million global users on March 2, 2012.

Narayanan at el. **??**

SNS Twitter Flickr Foursquare

Social structure data Profile data

Foursquare

3.1 Ground truth

To verify our de-anonymizing results , we have to determine the ground truth, i.e., the true mapping between the users of the online social networks. Actually, we do not need to label the mapping of all users since the ground truth as a test set can be far smaller than the complete network data.

[1]http://www.twitter.com
[2]http://www.flickr.com
[3]http://www.foursquare.com

Instead of labeling the user mapping by human editors, there are several sources to get the ground truth.

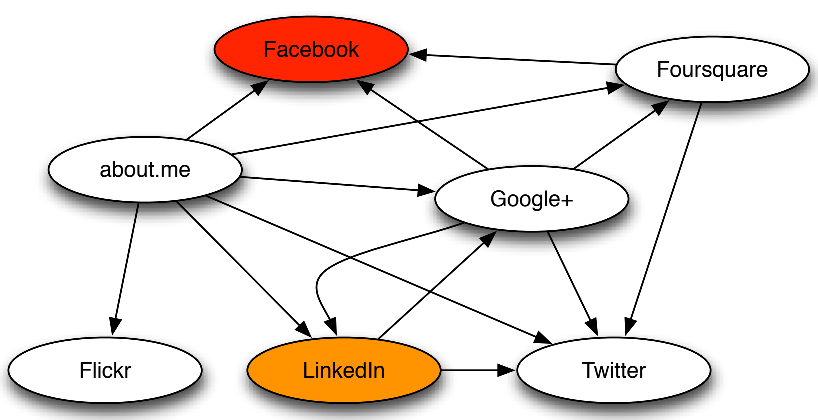
3.1.1 Single-source ground truth

About.me[4] is a personal web hosting service, which had at least 1 million users on October, 2011 [5]. The site offers registered users a simple platform from which to link multiple online identities, relevant external sites, and popular social networking websites such as Google+, Twitter, Facebook, LinkedIn, Flickr, YouTube, Foursquare. These links on user profile is naturally human-labelled mapping by the user itself, which can be seen as a zero-error ground truth. We picked a random sample of the mappings and verified by human inspection that the most of about.me users have Twitter accounts and at least one of Flickr and Foursquare accounts. About.me also provides simple APIs to list user directory and view the links on user profile without the strict crawling limitation. Therefore, we will mainly adopt the data from about.me to be our ground truth in this project.

3.1.2 Inferred multiple-source Ground truth

The links of the user profile page of the social networking websites are another great sources for ground truth, which is also generated by the user itself. Usually, a single user has many accounts for different social networking website. On the user profile page, there might be links to this user's accounts in the other popular social networking webiste. Especially, nowadays, for the most the social networking website, the user logs in with the connection to his/her Twitter or Facebook account, and that website may show the user's Twitter and Facebook account in the user profile. For example, the figure 1 shows how the links connects to other social networking website on the user profile page among the famous large social networking website: LinkedIn publicly shows the users' linked Twitter account and Gmail/Google+ account; and public Google+ profile reveals the user's Facebook and Twitter account; and Foursquare will show the user's login Twitter or Facebook account information.

Figure 1: Links on the user profile page of serveral social networking website



Fortunately, on these famous social networking webiste in the figure 1, the most of user's profile

[4]http://about.me

[5]http://techcrunch.com/2011/10/17/about-mes-ceo-on-how-to-hit-a-million-users-in-300-days-figure-out-who-your-entourage-is/

pages are publicly accessbile. A crawler can easily follow these links on the profile page, discover all linked accounts about one user, and even retrieve the user's real name and affiliation from the profile on the real-name social networking website, such as LinkedIn and Facebook (colorred in orange and red in figure 1. Thus, we can build a ground truth by exploring all linked accounts of each user.

3.2 Evaluation

Since we have a ground truth, correct matches

We will compare Network Alignment [24] and Simulated Annealling [19].

Belief Propagation

4 Deliverable

We will Ultimately, we hope to create a

nal report which summarizes the various corre- lations between sentiment content and popularity, which also provides a detailed picture of the way in which quote variants with the same informational content vary over time as function of sentiment. Central to our project is the examination of the causal eect of sentiment upon popularity and we want to build a story which best tests this theory. On a more conceptual level, we would like to use these results to comment on the fundamental nature of the news; shifting the focus away from its traditional conception as either an objective or even manipulative source of information, to more of a community of self-replicating entities whose properties arise out of internal features of sentiment, such as emotional variation and selection.

References

[1] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore Art Thou R3579X ? Anonymized Social Networks , Hidden Patterns , and Structural Steganography. 2007.

[2] M. Balduzzi, C. Platzer, and T. Holz. Abusing social networks for automated user profiling. *Recent Advances in ...*, pages 422–441, 2010.

[3] M. Bayati, M. Gerritsen, D. F. Gleich, A. Saberi, and Y. Wang. Algorithms for Large, Sparse Network Alignment Problems. *2009 Ninth IEEE International Conference on Data Mining*, 0:12, 2009.

[4] M. Bayati, D. Gleich, M. Gerritsen, and A. Saberi. Our motivation.

[5] M. Bayati, D. Gleich, A. Saberi, and Y. Wang. Message passing algorithms for sparse network alignment. *arXiv preprint arXiv:0907.3338*, 2009.

[6] S. Bradde, a. Braunstein, H. Mahmoudi, F. Tria, M. Weigt, and R. Zecchina. Aligning graphs and finding substructures by a cavity approach. *EPL (Europhysics Letters)*, 89(3):37009, Feb. 2010.

[7] M. Burkhart, D. Schatzmann, B. Trammell, E. Boschi, and B. Plattner. The role of network trace anonymization under attack. *ACM SIGCOMM Computer Communication Review*, 40(1):5, Jan. 2010.

[8] F. Cromi. An Alignment Algorithm using Belief Propagation and a Structure-Based Distortion Model. (April):166–174, 2009.

[9] A. L. Delcher, A. Phillippy, J. Carlton, and S. L. Salzberg. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research*, 30(11):2478–2483, 2002.

[10] X. Ding, L. Zhang, Z. Wan, and M. Gu. A Brief Survey on De-anonymization Attacks in Online Social Networks. *2010 International Conference on Computational Aspects of Social Networks*, pages 611–615, Sept. 2010.

[11] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Ontology Matching : A Machine Learning Approach. pages 1–20.

[12] M. El-Kebir, J. Heringa, and G. W. Klau. Lagrangian Relaxation Applied to Sparse Global Network Alignment. *Life Sciences*, 7036:225–236, 2011.

[13] J. Flannick, A. Novak, C. B. Do, B. S. Srinivasan, and S. Batzoglou. Automatic parameter learning for multiple local network alignment. *Journal of computational biology : a journal of computational molecular cell biology*, 16(8):1001–22, Aug. 2009.

[14] G. W. Klau. A new graph-based method for pairwise global network alignment. *BMC Bioinformatics*, 10(Suppl 1):S59, 2009.

[15] G. Kollias, S. Mohammadi, and A. Grama. Network Similarity Decomposition (NSD): A Fast and Scalable Approach to Network Alignment. *IEEE Transactions on Knowledge and Data Engineering*, PP(January):1, 2011.

[16] G. Kollias, M. Sathe, O. Schenk, and A. Grama. Fast Parallel Algorithms for Graph Similarity and Matching. *docs.lib.purdue.edu*, 2012.

[17] D. Koutra, A. Parikh, A. Ramdas, and J. Xiang. Algorithms for Graph Similarity and Subgraph Matching. 2011.

[18] M. Koyutürk, Y. Kim, U. Topkara, S. Subramaniam, W. Szpankowski, and A. Grama. Pairwise alignment of protein interaction networks. *Journal of computational biology : a journal of computational molecular cell biology*, 13(2):182–99, Mar. 2006.

[19] P. Kreitmann. CS224W: Project Writeup. pages 1–12, 2011.

[20] O. Kuchaiev. Global Network Alignment. pages 1–8, 2007.

[21] C.-S. Liao, K. Lu, M. Baym, R. Singh, and B. Berger. IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 25(12):i253–i258, 2009.

[22] V. Memišević and N. Pržulj. C-GRAAL: common-neighbors-based global GRAph ALignment of biological networks. *Integrative biology : quantitative biosciences from nano to macro*, 4(7):734–43, July 2012.

[23] S. Mohammadi and A. Grama. *Biological Network Alignment*.

[24] A. Narayanan and V. Shmatikov. Robust De-anonymization of Large Sparse Datasets. *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125, May 2008.

[25] A. Narayanan and V. Shmatikov. De-anonymizing Social Networks. *2009 30th IEEE Symposium on Security and Privacy*, pages 173–187, May 2009.

[26] R. a. Pache, A. Céol, and P. Aloy. NetAligner–a network alignment server to compare complexes, pathways and whole interactomes. *Nucleic acids research*, 40(Web Server issue):W157–61, July 2012.

[27] R. A. Pache, A. Céol, and P. Aloy. NetAligner–a network alignment server to compare complexes, pathways and whole interactomes. *Nucleic acids research*, 40(Web Server issue):W157–61, July 2012.

[28] P. Peng. A Local Algorithm for Finding Dense Bipartite-Like Subgraphs. *Computing and Combinatorics*, pages 145–156, 2012.

[29] E. Shi and B. I. P. Rubinstein. Link Prediction by De-anonymization: How We Won the Kaggle Social Network Challenge.

[30] A. Todor, A. Dobra, and T. Kahveci. Probabilistic Biological Network Alignment. 6(1):1–14, 2007.

[31] Y. Wang. A Genetic Algorithm and its Parallelization for Graph Matching with Similarity Measures Department of Intelligence and Computer Science Nagoya Institute of Technology.

[32] Y. Wang, N. Ishii, and C. Science. Graph Matching, Similarity Measures, Genetic Algorithms, Parallel Computing, similarity- based approximate reasoning. 1. (2):2–5.

[33] G. Wondracek and T. Holz. A practical attack to de-anonymize social network users. *Security and Privacy (SP . . .*, 2010.

5 Appendix