

# CS224W Course Project Proposal

## De-anonymizing social networks via network alignment

Danqi Chen  
Stanford University  
danqi@stanford.edu

Botao Hu  
Stanford University  
botaohu@stanford.edu

Shuo Xie  
Stanford University  
shuoxie@stanford.edu

October 23, 2012

### 0.1 Problem Formulation

Formally, the problem of de-anonymizing social networks is to identify the same users between two anonymized social networks. A social network can be modeled as a directed graph (e.g., Twitter, Flickr) or an undirected graph (e.g., Facebook): Each node in the graph represents a user in the social network and each edge between two nodes characterizes the relationship between the users, such as “following” in Twitter or “friendship” in Facebook. In addition, anonymized graphs are usually released with at least some attributes in their nodes such as user profile information, user interests, corresponding to a set of attributes for each node in  $V$ , which may help us making de-anonymization easier.

## 1 Literature Review

### 1.1 De-anonymizing Social Networks

[25] is the first work to perform large-scale de-anonymization of real-world social networks. Specifically, they try to match the users between Titter and Flickr and show that 30.8% of the verifiable members of both social networks could be recognized with 12% error rate. In their paper, they propose a greedy-based de-anonymization algorithm, which consists of two steps - “seed finding” and “propagation”. In the “seed finding” step, they identify a small number of “seed” nodes which are present in the both networks and map them to each other; Later they use the seed nodes as “anchors” to propagate the de-anonymization to more and more nodes. They iteratively try to extend the mapping as follows: pick an arbitrary unmapped node in a graph, and map it to the “most similar” node in the other graph, where the similarity score is computed based on some heuristics such as cosine similarity or the number of formed squares.

It is worth noting that [?] won the Kaggle Social Network Challenge by applying the above de-anonymizing algorithm to the contest dataset and real Flickr network, and utilize the connectivity among “matched users” in Flickr to perform link predictions.

Although the proposed algorithm demonstrates the feasibility of successful de-anonymization on large-scale social networks, the algorithm is based on simple greedy approaches, so that the result is far from “perfect matchings”; Furthermore, it is purely based on the network topology and doesn’t incorporate any attribute of nodes, and these public attributes may greatly improve the de-anonymization process.

## 1.2 Network Alignment

Network alignment is another well-formed problem, which aims to perform matchings or alignments between the vertices of two undirected graphs. This problem has been studied extensively, and has been applied successfully in many domains such as finding common pathways in biological networks([?, ?, 21]), ontology matchings between the Library of Congress with the categories from Wikipedia([3]), and multi-lingual ontologies([19, 5]). In this subsection, we will review the existing algorithms that produce good solutions for the network alignment, analyze the strengths and weakness of these algorithms and finally investigate the possibility to apply the network alignment algorithms on our settings of de-anonymizing social networks.

Most existing work formulate the network alignment problem as follows: given two sets of vertices  $V_A = \{1, 2, \dots, n\}$ ,  $V_B = \{1', 2', \dots, m'\}$ , and  $A = (V_A, E_A)$ ,  $B = (V_B, E_B)$  are two undirected graphs with their respective vertex and edge sets. Let  $L$  be a bipartite graph between the vertices of  $A$  and  $B$ , formally  $L = (V_A \cup V_B, E_L)$  (in many algorithm settings,  $L$  is the complete bipartite graph). The goal is to find a matching  $M$  between  $V_A$  and  $V_B$  using only edges from  $L$  ( $M \subseteq L$ ), such that the number of overlapped edges is maximized. Here an edge  $(i, j) \in E_A$  is *overlapped* iff.  $(i', j') \in E_B$ ,  $(i, i')$  and  $(j, j')$  belong to  $M$ .

In a more general setting following [?], each edge  $e \in E_L$  has a non-negative weight  $w_e$ , and the goal is to find a matching  $M$  maximizing a linear combination of the matching weight and the number of over-lapped edges. Mathematically, it can be formulated as a quadratic program(QP): define  $\mathbf{S}$  as a 0-1 matrix of size  $|E_L| \times |E_L|$  where  $S[ii', jj'] = 1$  if  $(i, j) \in E_A$  and  $(i', j') \in E_B$ ,  $\mathbf{A}$  as the binary incidence matrix of  $L$  with dimensions  $|V_L| \times |E_L|$ . We want to find a 0-1 vector  $\mathbf{x}$  such that:

$$\begin{aligned} \max_{\mathbf{x}} \quad & \alpha \mathbf{w}^\top \mathbf{x} + \beta \mathbf{x}^\top \mathbf{S} \mathbf{x} \\ \text{subject to} \quad & \mathbf{A} \mathbf{x} \leq \mathbf{1}, x_{ii'} \in \{0, 1\} \end{aligned}$$

Solving the QP formulation is NP-hard, therefore many existing works have been attempted to relax the constraints or find the heuristics:

**IsoRank**([?]) learns a similarity measure  $r_{ii'}$  between node  $i$  in  $V_A$  and node  $i'$  in  $V_B$  and  $r_{ii'}$  is defined as:

$$r_{ii'} = \sum_{(i,j) \in E_A} \sum_{(i',j') \in E_B} \frac{R_{j,j'}}{|N_j||N_{j'}|}.$$

The new weights  $r_{ii'}$  can be found using an eigen-value calculation as in PageRank. Later, [5] proposes an extension of IsoRank algorithm based on the Kronecker product, named **SpaIsoRank**, which can be adapted to the case when  $L$  is sparse. [15] proposes a **Network Similarity Decomposition (NSD)** algorithm, which uses matrix decomposition to approximate the power iteration in IsoRank and runs much faster than the original algorithm.

[14] formulates a **linear programming(LP)** relaxation of the original QP, and proposes an iterative approximation algorithm, which is a tighter LP relaxation and Lagrangian decomposition of the symmetry constraints. This algorithm can be used to handle large networks when  $L$  is sparse.

[5] proposes two message passing algorithms **MP** and **MP++** for network alignment: It constructs a fractor graph and defines a probability distribution on the space of all matchings in  $L$  that assigns the highest probability to the matching that maximize the QP, and then applies the stanford Belief Propagation algorithm for finding the optimum solution. As well, MP algorithms can be applied to large sparse networks.

In last year’s course project [19], Kreitmann proposes a simple **simulated annealing(SA)** algorithm, which initializes a greedy-based matching first, and adopts simulated algorithm approach taking the number of “overlapped” edges as the heuristic function for swapping two matches.

Although the above algorithms may produce good results in their applications, we still have some concerns:

- To our knowledge, no network alignment algorithm has been applied to the problem of de-anonymizing social network.
- IsoRank, LP, and MP algorithms can only be applied in small datasets (e.g., biological networks) or large-scale sparse networks. For handling large-scale networks, how to select  $L$  should be the issue we need to consider.
- All the network alignment algorithms are based only on the network structure, no node attribute is incorporated to improve the alignment.
- All the above algorithms are applying to aligning two undirected networks, we should consider if it is possible to develop network alignment algorithms for two directed network or one undirected network and one directed network, which can be adopted to our social networks.

## 2 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 accessible by crawlers or APIs.

The first graph is the “following” relationships on the Twitter<sup>1</sup>, a microblogging service, which has 500 million users (200 million active). We consider to adopt the data crawled by Kwak et al.<sup>2</sup> containing 41 million users. In order to increase the overlap to the other two social networks, we will extend this dataset to the latest network as possible.

The second graph is the “contact” relationship on Flickr<sup>3</sup>, 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<sup>4</sup>, a location-based social network, which has 22 million global users on March 2, 2012.

Narayanan et al. [24] did the experiment on aligning Twitter and Flickr data.

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

---

<sup>1</sup><http://www.twitter.com>

<sup>2</sup><http://an.kaist.ac.kr/traces/WWW2010.html>

<sup>3</sup><http://www.flickr.com>

<sup>4</sup><http://www.foursquare.com>

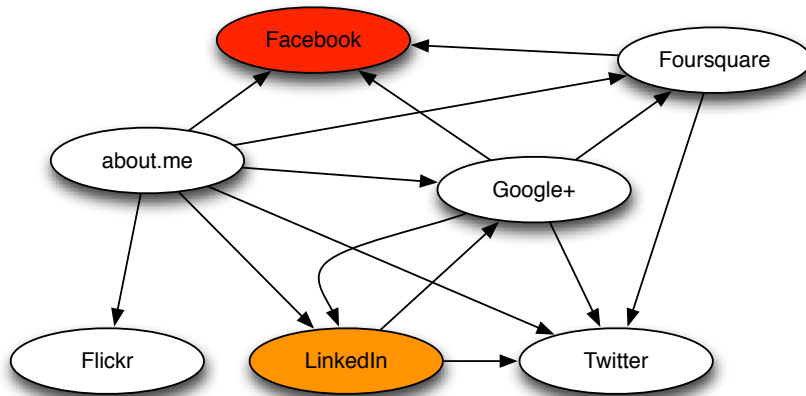
### 2.1.1 Single-source ground truth

About.me<sup>5</sup> is a personal web hosting service, which had at least 1 million users on October, 2011<sup>6</sup>. 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.

### 2.1.2 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 website. 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 website in the figure 1, the most of user's profile pages are publicly accessible. 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 (colored in

<sup>5</sup><http://about.me>

<sup>6</sup><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/>

orange and red in figure 1. Thus, we can build a ground truth by exploring all linked accounts of each user.

## 2.2 Evaluation

We will compare our algorithm on the real dataset to Network Alignment [24] and Simulated Annealing [19].

Accuracy: Given a ground truth, the accuracy evaluation can be simply the correct matches between two networks.

Scalability: Running time on the large-scale data.

## 3 Deliverable

We will implement codes in SNAP<sup>7</sup> framework and integrate the complete component of network alignment into the SNAP package.

## 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.

---

<sup>7</sup><http://snap.stanford.edu>

- [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.

## 4 Appendix