An Efficient Reconciliation Algorithm for Social Networks

R. Collin Guieb – University of California, Los Angeles

ABSTRACT

In this digital age, social networks play a major role in the way we interact as human beings. From Facebook and Twitter to LinkedIn and Google+, people have so many different accounts that it is often difficult to keep track of all our relationships and connections over so many different websites. The goal of a recent study done by Nitish Korula and Silvio Lattanzi was to create an algorithm to reconcile an individual's networks by identifying all the accounts belonging to the same individuals. Using social networks as their models for their experiment, the two were able to successfully develop a simple, efficient algorithm to solve this social network reconciliation problem.

1. INTRODUCTION

Social networks allow us to share and exchange information; connect and network with our peers; and most importantly, unite together as a human race. With the rise of the digital age, online social networking has become more and more ingrained into our society. For example, Facebook with its 1.44 billion monthly active users (as of April 2015) outnumbers the population of China, effectively making Facebook the largest country in the world [2].

With such a huge reliance on social media and so many active social networkers, it becomes hard to keep track of all of one's connections and relationships. Recently, a team of Google researchers proposed to find a way of reconciling users' online relationships with their real life relationships. In order to accomplish this, they created a model consisting of various social networks and developed an algorithm to identify individuals across these networks and reconcile their accounts. Their goal was to create an algorithm that was simple, parallelizable, and robust to malicious users or fake accounts.

2. MODEL

Since each social network serves a different purpose, users often have different networks on each individual site. For example, a network of Facebook friends usually represents personal and familial connections. In contrast, a network of LinkedIn users embodies professional and work relationships. For these reasons, our model needs to include as many social networks identify all the accounts belonging to a to the same individual across different networks.

Our goal is to find a model of the "true" underlying social network *G*(*V, E*), where each person is a vertex *V* and each relationship or connection between two people is an edge *E*. We will attempt to find this true network of relationships by consolidating and reconciling *n* online social networks *G1*(*V1, E1*), *G2*(*V2, E2*) ... *Gn*(*Vn, En*). Essentially, we want to use social networking sites to create an isomorphic graph to the actual graph of real world relationships.

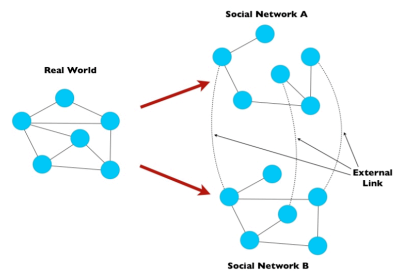

Figure 1: A simple diagram of the study [1]. Our goal is to use our algorithm to obtain a model of real world networks using Social Network A and Social Network B.

It is important to note that this process would not be extremely difficult unless we already have existing external links between networks. Fortunately, a small fraction of users already connects their accounts. For example, someone may post a link to their Twitter account on their Facebook or Google+ or vice versa.

3. ALGORITHM

Simply put, the first step of the algorithm was to match users from Social Network A to Social Network B. This was done using the User-Matching algorithm as follows:

Input:
$G_1(V, E_1), G_2(V, E_2), L$ a set of initial identification links across the networks, D the maximum degree in the graph a minimum matching score T and a specified number of iteration k.
Output:
A larger set of identification links across the networks.
Algorithm:
For $i = 1, \dots, k$
  For $j = \log D, \dots, 1$
    For all the pairs (u, v) with $u \in G_1$ and $v \in G_2$ and such that $d_{G_1}(u) \geq 2^j$ and $d_{G_2}(v) \geq 2^j$
      Assign to (u, v) a score equal to the number of similarity witnesses between u and v
    If (u, v) is the pair with highest score in which either u or v appear and the score is above T add (u, v) to L.
Output L

Essentially, the vertices of the graph of Social Network A were iterated over and compared to vertices of Social Network B. Each comparison was assigned a similarity score. If the similarity score reached a certain threshold value *T*, then the two vertices were expected to be the same user.

After the user-matching process, the duplicate vertices from Social Network B that matched Social Network A were deleted, and the edges were consolidated. This overall process was repeated for *n* social networks. The higher the *n,* the closer we simulate our "true" social network of the real world.

4. EXPERIMENT & RESULTS

A series of different experiments were run using the algorithm described above.

Network	Number of nodes	Number of edges
PA [5]	1,000,000	20,000,000
RMAT24 [7]	8,871,645	520,757,402
RMAT26 [7]	32,803,311	2,103,850,648
RMAT28 [7]	121,228,778	8,472,338,793
AN [19]	60,026	8,069,546
Facebook [30]	63,731	1,545,686
DBLP [1]	4,388,906	2,778,941
Enron [16]	36,692	367,662
Gowalla [8]	196,591	950,327
French Wikipedia [2]	4,362,736	141,311,515
German Wikipedia [2]	2,851,252	81,467,497

Table 1: The datasets used for each experiment [1].

In summary, the experiments ran on each of the different networks were successful. Errors where the user-matching algorithm incorrectly matched users produced approximately 5% errors overall. As expected, the results with higher values of threshold *T* returned less error.

In addition to social networks, the researchers tested the parallelizability of the algorithm by trying to reconciling the French and German Wikipedia databases. While the algorithm was able to work on these databases, the results were far less accurate than the experiments done on the social network databases. This fact is mainly due to language barriers and human error when reconciling the article links. When dealing with languages, the two networks will be similar, but they do not necessarily share a single common source.

5. CONCLUSION

Overall, the authors of the study were able to successfully develop and implement an algorithm for social network reconciliation. The algorithm was also somewhat parallel and come be used on other similar databases and networks, however, to a slightly lesser degree of success.

This study has several practical applications that could be useful for everyday life. The results from this experiment provide us with information about users and their social network habits. From this, we get a better of social dynamics and how people interact online in comparison to how people interact in the physical world.

Additionally, by finding information across different social networks, we can develop and improve more personalized content. By analyzing a user's connections on other social networks, Social Network A can provide more personalized "who to follow" or "friend suggestion" lists.

If done on a large enough scale, this algorithm could potentially provide a worldwide phonebook including everyone's different social media accounts and who they know. However, while this may be a breakthrough in technology and may seem extremely practical, it raises some serious privacy issues. People usually do not link their accounts for a reason, and most likely would not want one giant social network for everyone to access.

REFERENCES

[1] Korula, Nitish, and Silvio Lattanzi. "An Efficient Reconciliation Algorithm for Social Networks." *Proc. VLDB Endow. Proceedings of the VLDB Endowment* 7.5 (2014): 377-88. VLDB, Jan. 2015. Web. 4 Dec. 2015.

[2] Hatmaker, Taylor. "Facebook Officially Becomes World's Largest Country." *The Daily Dot*. Daily Dot Media, 22 Apr. 2015. Web. 04 Dec. 2015.

[3] Malhotra, Anshu, Luam Totti, Wagner Meira, Jr., Ponnurangam Kumaraguru, and Virgilio Almeida. "Studying User Footprints in Different Online Social Networks." *Studying User Footprints in Different Online Social Networks*. Proceeding ASONAM '12 Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining, 2012. Web. 04 Dec. 2015.