An Efficient Reconciliation Algorithm for Social Networks

Collin Guieb
CS35L Lab 3

OUTLINE

- Introduction
- Model
- Algorithm
- Experiment
- Conclusions

INTRODUCTION: Context

- In the 21st century, social life is online
- Connects people and stimulates interaction via the internet

INTRODUCTION: The Problem

- Social Network Reconciliation Problem:
  - Each network represents a subset of your “real” ego-network
  - We want to identify you across all social media platforms

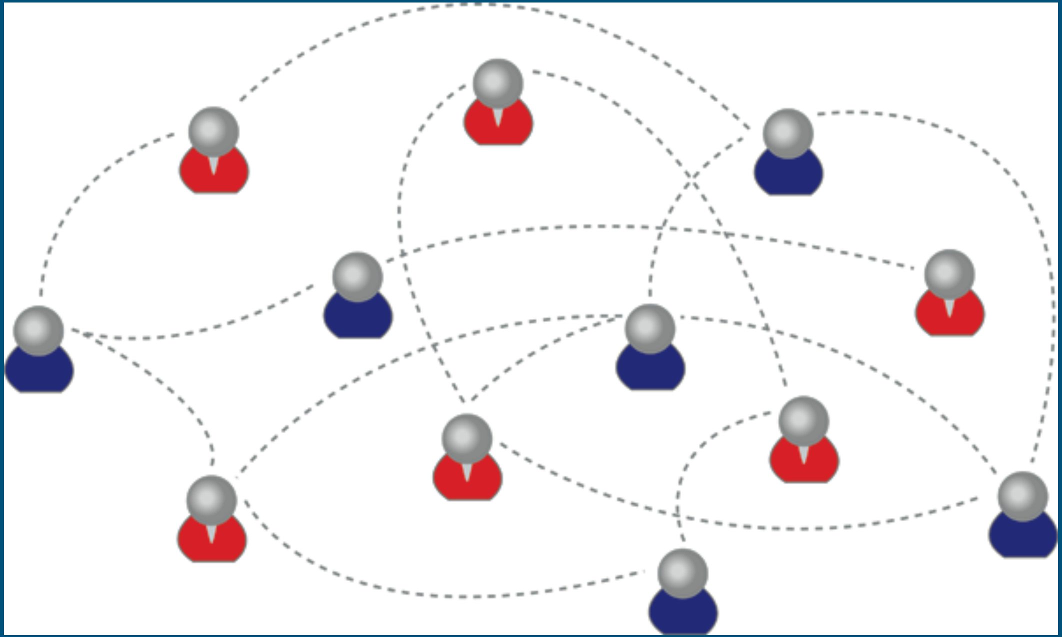
The Conversation
The Art of Listening, Learning, and Sharing
Brought to you by
Brian Solis and JESS3
Social Bookmarks
Comment & Reputation
Crowdsourced Content
Blog Platforms
Blogs/Conversations
Blog Communities
Micromedia
Lifestreams
Specific to Twitter
SMS/Voice
Social Networks
Niche Networks
Customers Service Networks
Location
Video
Video Aggregation
Documents
Events
Music
Wiki
LiveCasting - Video and Audio
Pictures

INTRODUCTION: Goals

- We want to create an algorithm that is:
  - simple
  - parallelizable
  - robust to malicious users

MODEL

- BASIC GRAPH THEORY
- Ideally, we want to create a "true" unifying graph G (V, E).
- Start with G1(v1, e1), G2(v2, e2), … Gn(vn, en)



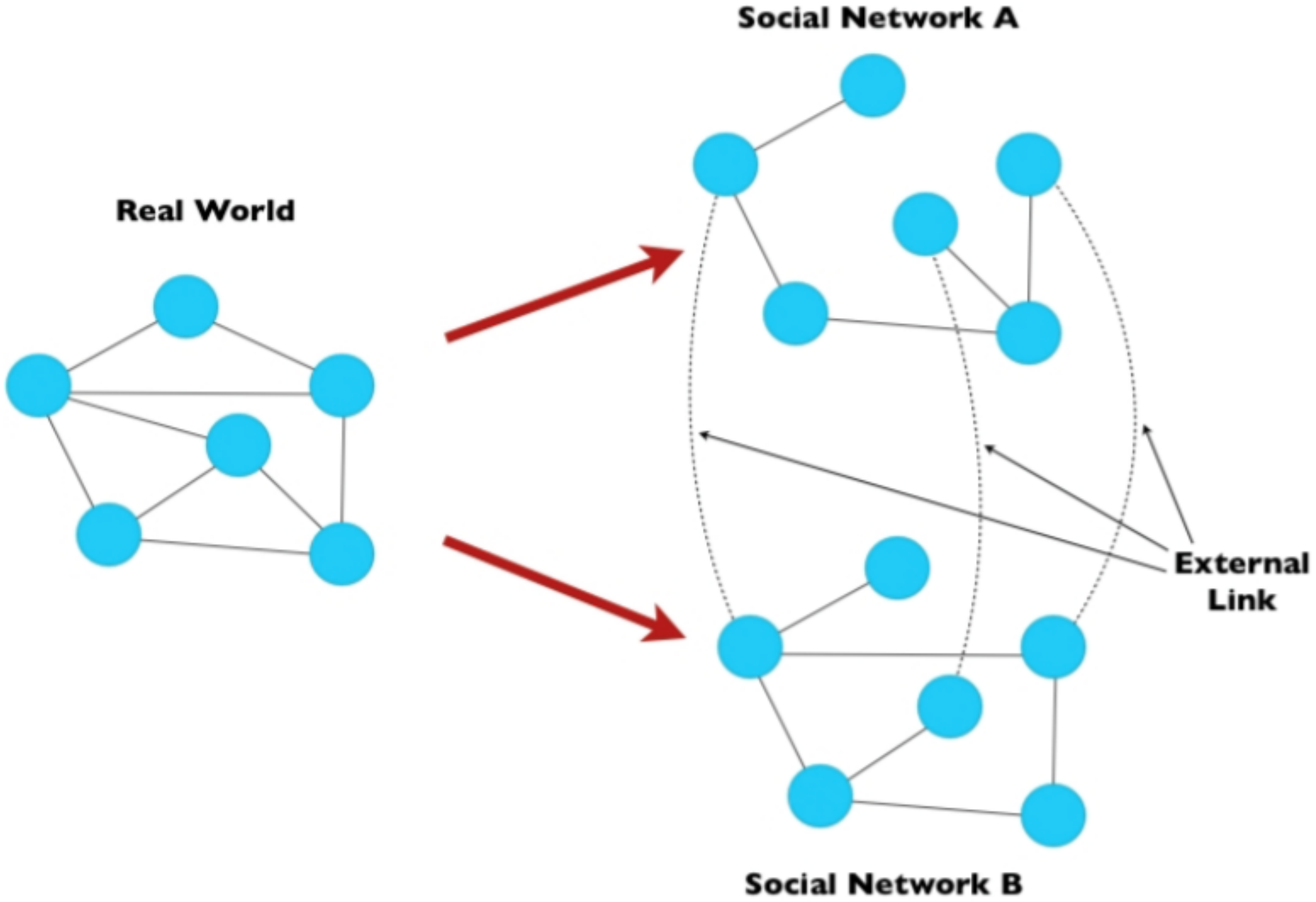
ALGORITHM

1. Match users from social network G1(v1,e1) to social network G2(v2,e2) using the User-Matching algorithm.

2. Delete duplicates and consolidate edges.

3. Repeat process for *n* social networks.

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 $i = 1, \ldots, k$
  For $j = \log D, \ldots, 1$
    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


Real World
Social Network A
External
Link
Social Network B

EXPERIMENT

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

EXPERIMENT

Pr	Threshold 5		Threshold 4		Threshold 2	
	Good	Bad	Good	Bad	Good	Bad
20%	23915	0	28527	53	41472	203
10%	23832	49	32105	112	38752	213
5%	11091	43	28602	118	36484	236

Pr	Threshold 5		Threshold 4		Threshold 3	
	Good	Bad	Good	Bad	Good	Bad
10%	3426	61	3549	90	3666	149

TABLE 1: Results for Facebook (top) and Enron (bottom).

EXPERIMENT

Pr	Threshold 5		Threshold 3	
	Good	Bad	Good	Bad
10	108343	9441	122740	14373

TABLE 2: Combined results for the French and German Wikipedia pages.

CONCLUSIONS

- Practical Applications:
  - better understanding of social dynamics
  - improved personalized content
  - essentially creating a worldwide phonebook
- Algorithm can be used on other similar databases
- Privacy Issues?