Applications of Social Network Analysis to Online Social Networks

Thomas Breuel

Exploring the Google+ Social Graph

Magno et. al, IMC 2012

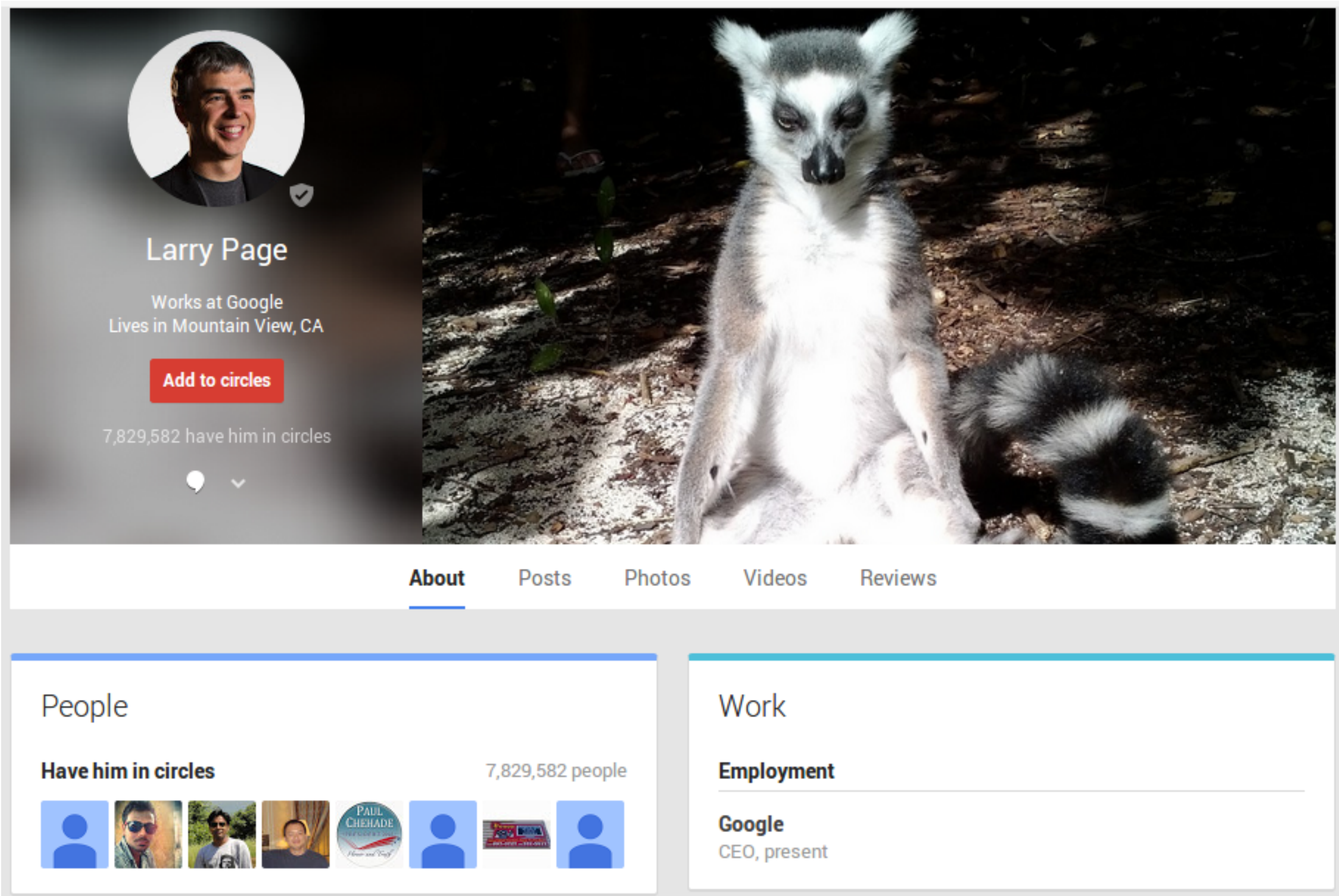
Google+

- users (linked to GMail accounts)
  - real names
  - user info (personal, professional, …)
  - interests
  - education

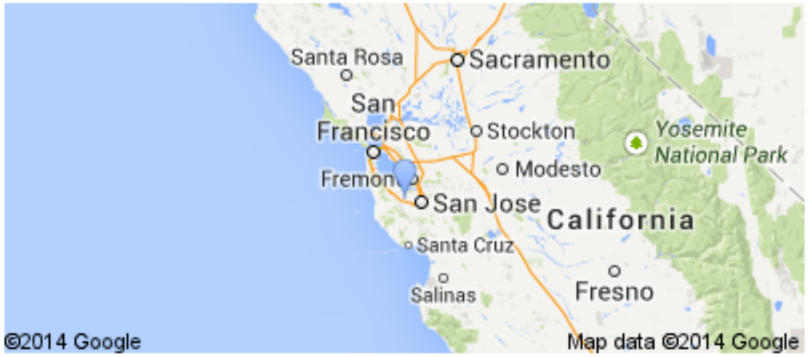
- links between users
  - directed relationship
  - users organized into circles
  - publish-to-circle
  - subscribe-to-all-circles (can narrow down)

Google+ main results

- 7 of the 20 top users are IT professionals
- 1% of users share contact information
- lots of single males share home phone
- sharing differs greatly between countries
- physical distance crucial in formation of links
- global and national links differ greatly by country


Larry Page
Works at Google
Lives in Mountain View, CA
Add to circles
7,829,582 have him in circles
About
Posts
Photos
Videos
Reviews
People
Have him in circles
7,829,582 people
Work
Employment
Google
CEO, present

Places



Currently
Mountain View, CA

Links

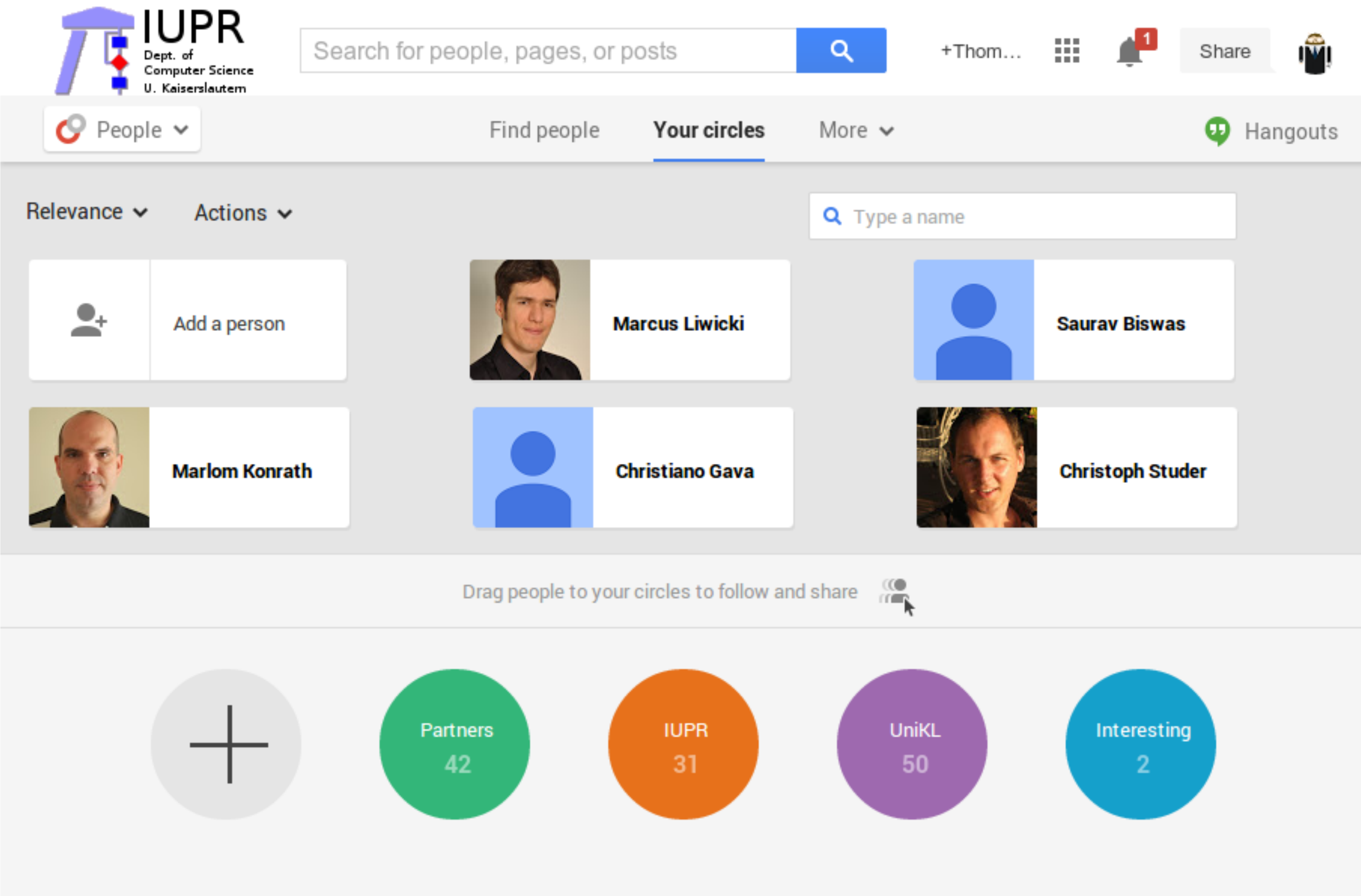
Google+ URL
google.com/+LarryPage

YouTube
Larry Page

Basic Information

Gender	Male
Relationship	Married

⊕ Add your own contact details about Larry. Visible only to you.


IUPR
Dept. of
Computer Science
U. Kaiserslautern
Search for people, pages, or posts
+Thom...
1
Share
People
Find people
Your circles
More
Hangouts
Relevance
Actions
Type a name
Add a person
Marcus Liwicki
Saurav Biswas
Marlom Konrath
Christiano Gava
Christoph Studer
Drag people to your circles to follow and share
Partners
42
IUPR
31
UniKL
50
Interesting
2

[Larry Page - Google+](https://plus.google.com/+LarryPage)

https://plus.google.com/+**LarryPage**

by Larry Page - in 7,755,216 Google+ circles

I have suggestions to you if u take it your business will increase ten folds for sure . first i will give my suggestions then implement it, if u get 2 times more profit just **...**

I'm excited to announce Calico ...
+Larry Page the difference between +Google and other **...**

Here is the speech I just gave ...
+Larry Page is still in the earnings call. Google Q2 2013 **...**

Larry Page - About
Larry Page. Works at Google. Lives in Mountain View, CA. 7 **...**

Larry Page - Videos
Larry Page - Google - Mountain View, CA. **...** Larry Page. Works **...**

More results from google.com »

Table 1: Top 20 users ranked by in-degree

Rank	Name	About
1	Larry Page	IT (Google)
2	Mark Zuckerberg	IT (Facebook)
3	Britney Spears	Musician
4	Snoop Dogg	Musician
5	Sergey Brin	IT (Google)
6	Tyra Banks	Model
7	Vic Gundotra	IT (Google)
8	Paris Hilton	Socialite
9	Richard Branson	Businessman (Virgin Group)
10	Dane Cook	Comedian
11	Jessi June	Model
12	Trey Ratcliff	Blogger
13	will.i.am	Musician
14	Felicia Day	Actor
15	Thomas Hawk	Blogger
16	Tom Anderson	IT (Myspace)
17	Pete Cashmore	IT (Mashable)
18	Guy Kawasaki	IT (Apple) & Writer
19	Wil Wheaton	Actor & Writer
20	Ron Garan	Astronaut (NASA)

Table 2: Public attributes available in Google+

Attribute	Available	%
Name	27,556,390	100.00
Gender	26,914,758	97.67
Education	7,471,191	27.11
Places lived	7,371,461	26.75
Employment	5,917,609	21.47
Phrase	4,075,132	14.79
Other profiles	3,713,546	13.48
Occupation	3,656,447	13.27
Contributor to	3,622,627	13.15
Introduction	2,149,191	7.80
Other names	1,210,760	4.39
Relationship	1,186,903	4.31
Braggin rights	1,074,964	3.90
Recommended links	1,001,349	3.63
Looking for	753,704	2.74
Work (contact)	60,434	0.22
Home (contact)	58,876	0.21

Table 3: **Information shared by all users and *tel-users***

	All users	Tel-users
Total	27,556,390	72,736
Gender (N)	26,914,758	71,267
Male	67.65%	85.99%
Female	31.46%	11.26%
Other	0.89%	2.75%
Relationship (N)	1,186,903	29,068
Single	42.82%	57.24%
Married	26.59%	21.03%
In a relationship	19.80%	10.23%
It's complicated	3.16%	3.98%
Engaged	4.39%	2.98%
In an open relationship	1.26%	2.77%
Widowed	0.50%	0.58%
In a domestic partnership	1.08%	0.77%
In a civil union	0.39%	0.41%
Location (N)	6,621,644	45,676
United States	31.38%	8.92%
India	16.71%	31.90%
Brazil	5.76%	4.72%
United Kingdom	3.35%	2.19%
Canada	2.30%	1.52%
Other	40.50%	50.77%

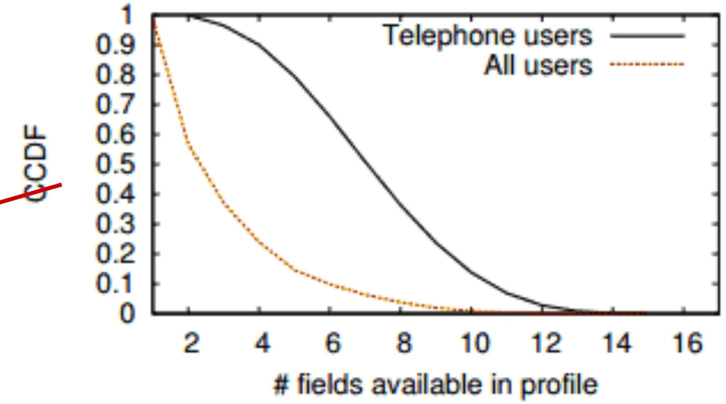


Figure 2: Number of fields shared by users in the profile

degree distribution

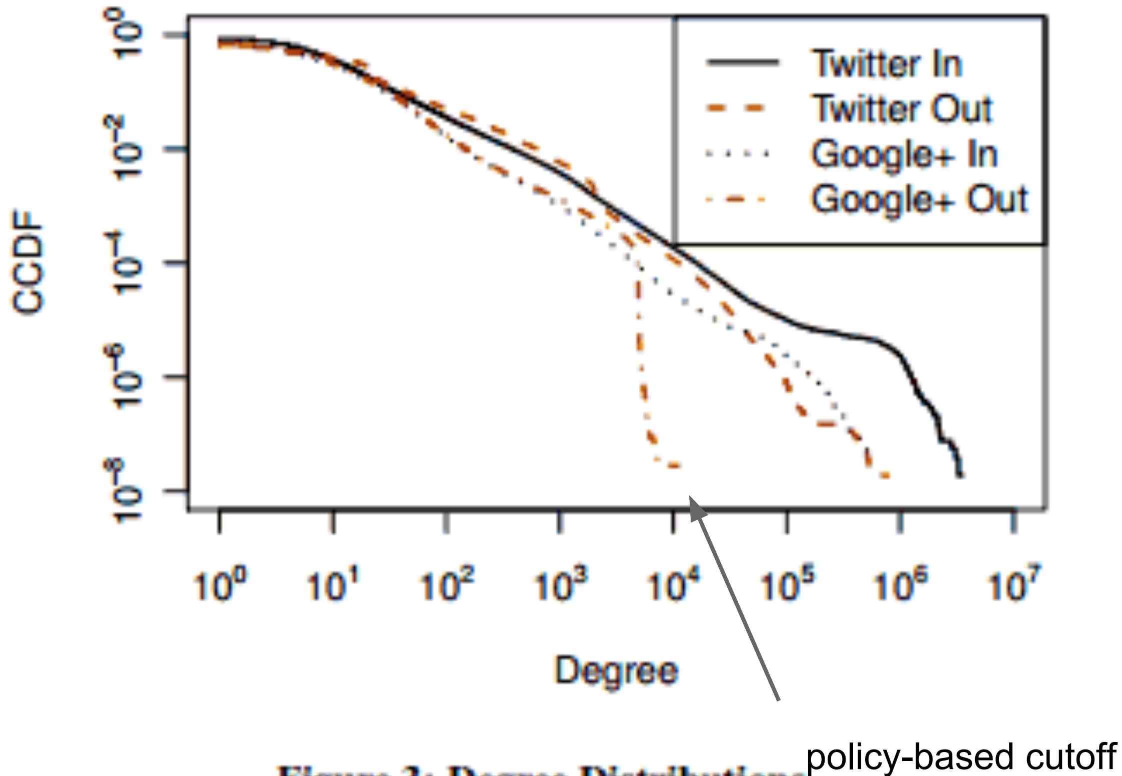
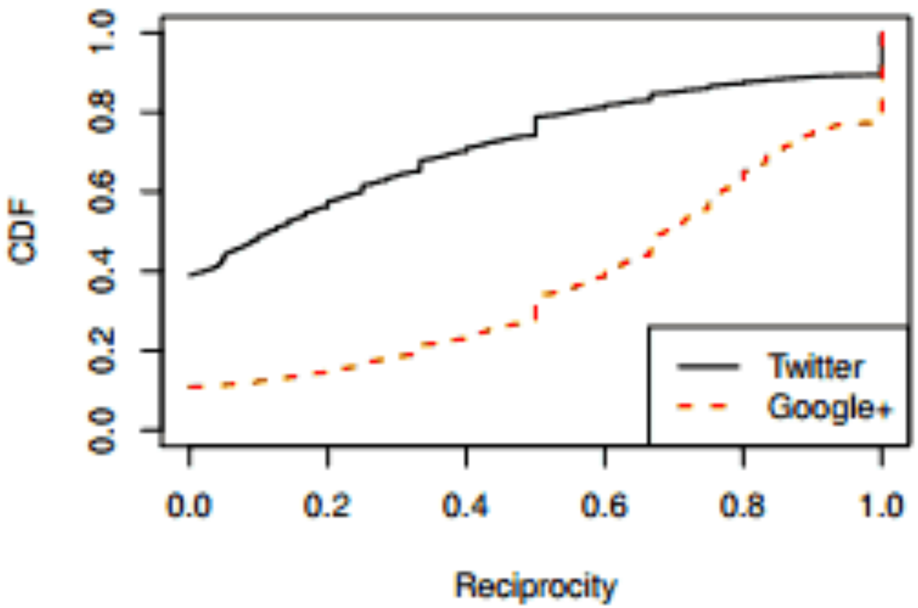


Figure 3: Degree Distributions

power law distribution typical of human social networks

reciprocity

$$RR(u) = \frac{|OS(u) \cap IS(u)|}{|OS(u)|}$$



(a) Reciprocal links

Which service has higher reciprocity?

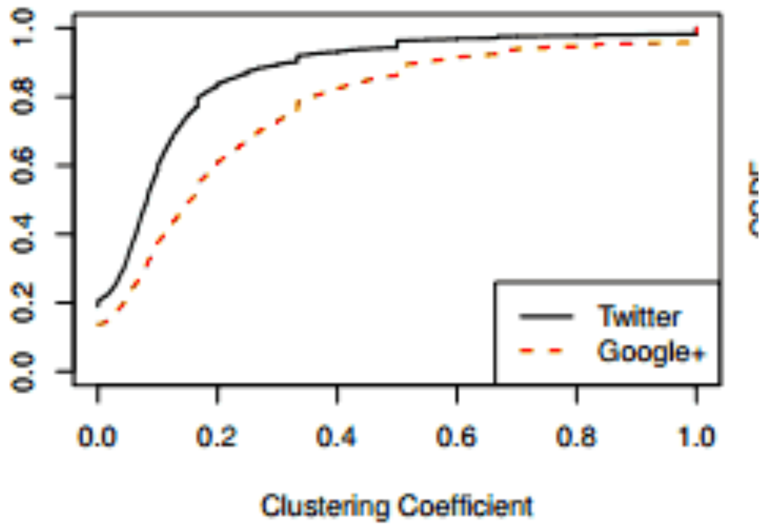
reciprocity

- Google+ has higher reciprocity of links than Twitter
- probable cause: fewer big media outlets on Google+

clustering coefficient

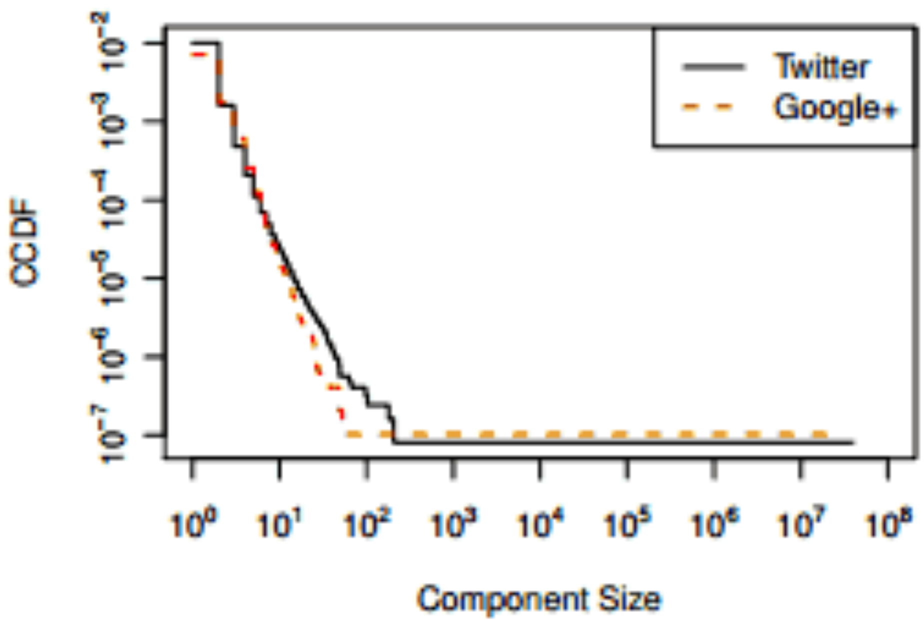
clustering = probability that two neighbors of a node are neighbors of each other (triangles)

Google+ has higher clustering coefficients (=more personal usage?)



strongly connected components

one weakly connected component from crawl
measure number and size of SCCs



(c) Size of the strongly connected components

degrees of separation

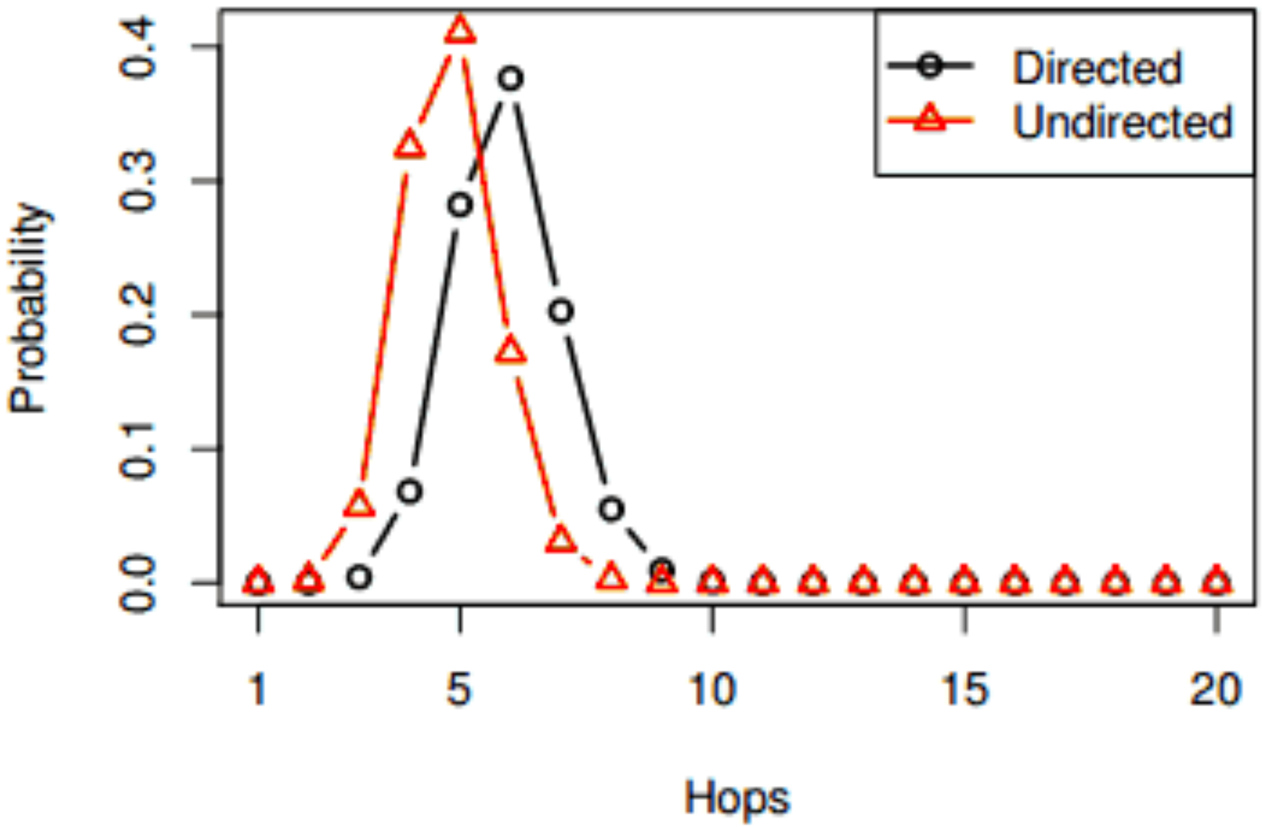


Figure 5: Estimated path length distribution

social graph statistics

Table 4: Comparison of topological characteristics of Google+ and other online social networks

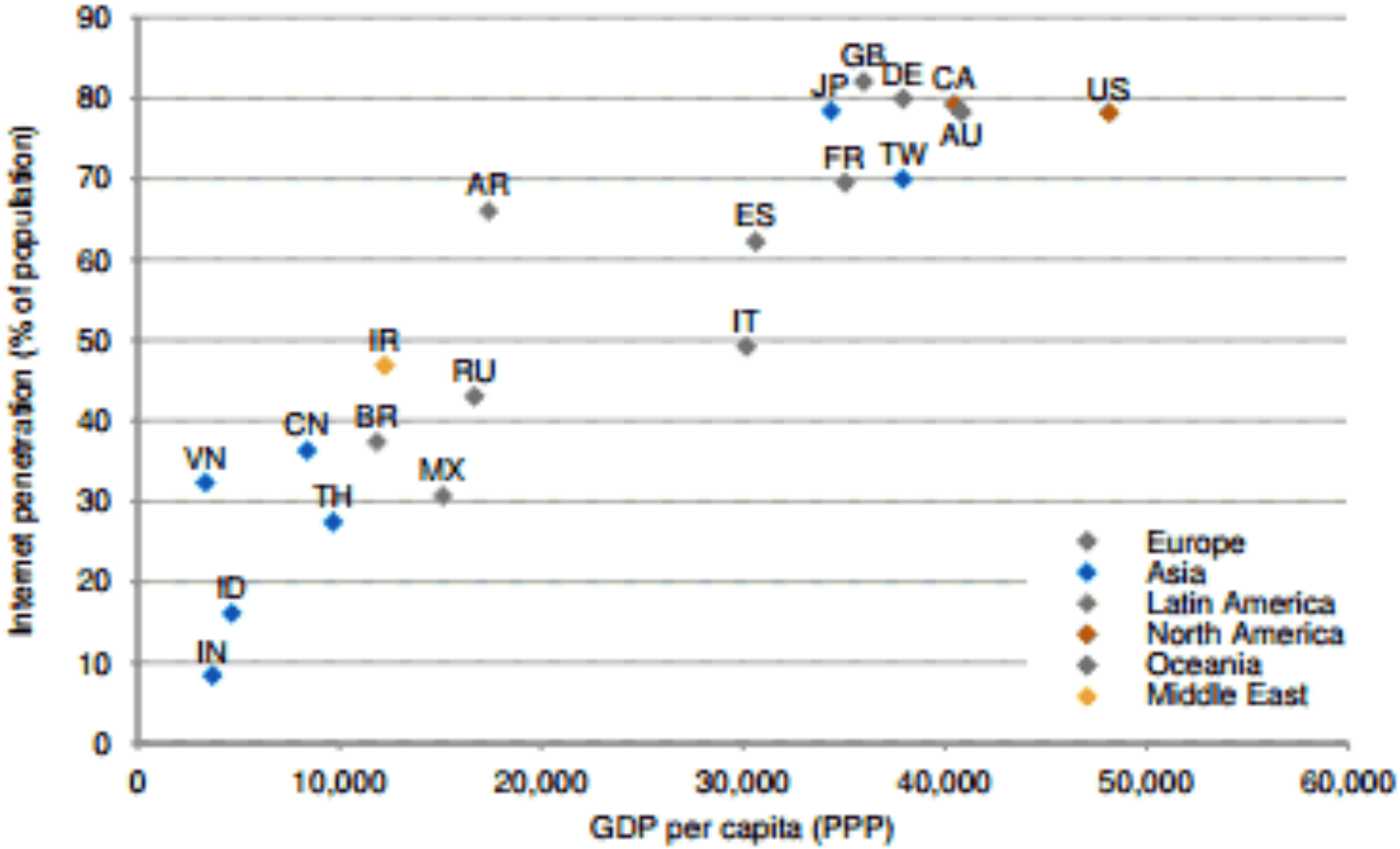
Network	Nodes	Edges	% Crawled	Path length	Reciprocity	Diameter	In-degree	Out-degree
Google+	35M	575M	56%	5.9	32%	19	16.4	16.4
Facebook	721M	62G	100%	4.7	100%	41	190.2	190.2
Twitter	41.7M	106M	100%	4.1	22%	18	28.19	29.34
Orkut	3M	223M	11%	4.3	100%	9	-	-

geographic quiz

put the following countries in descending order of per-capita GDP

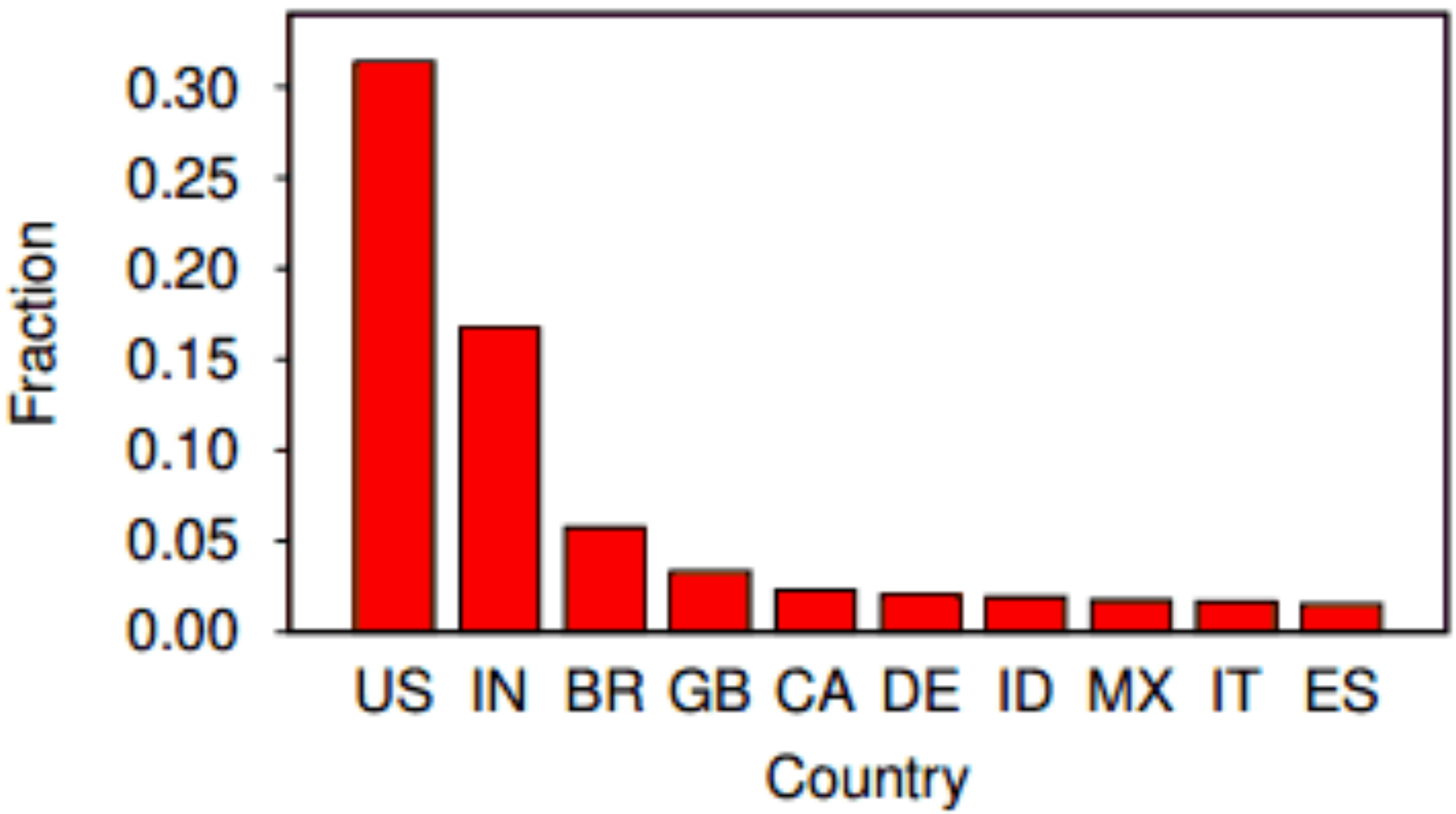
Russia, China, US, Germany, Japan

internet penetration

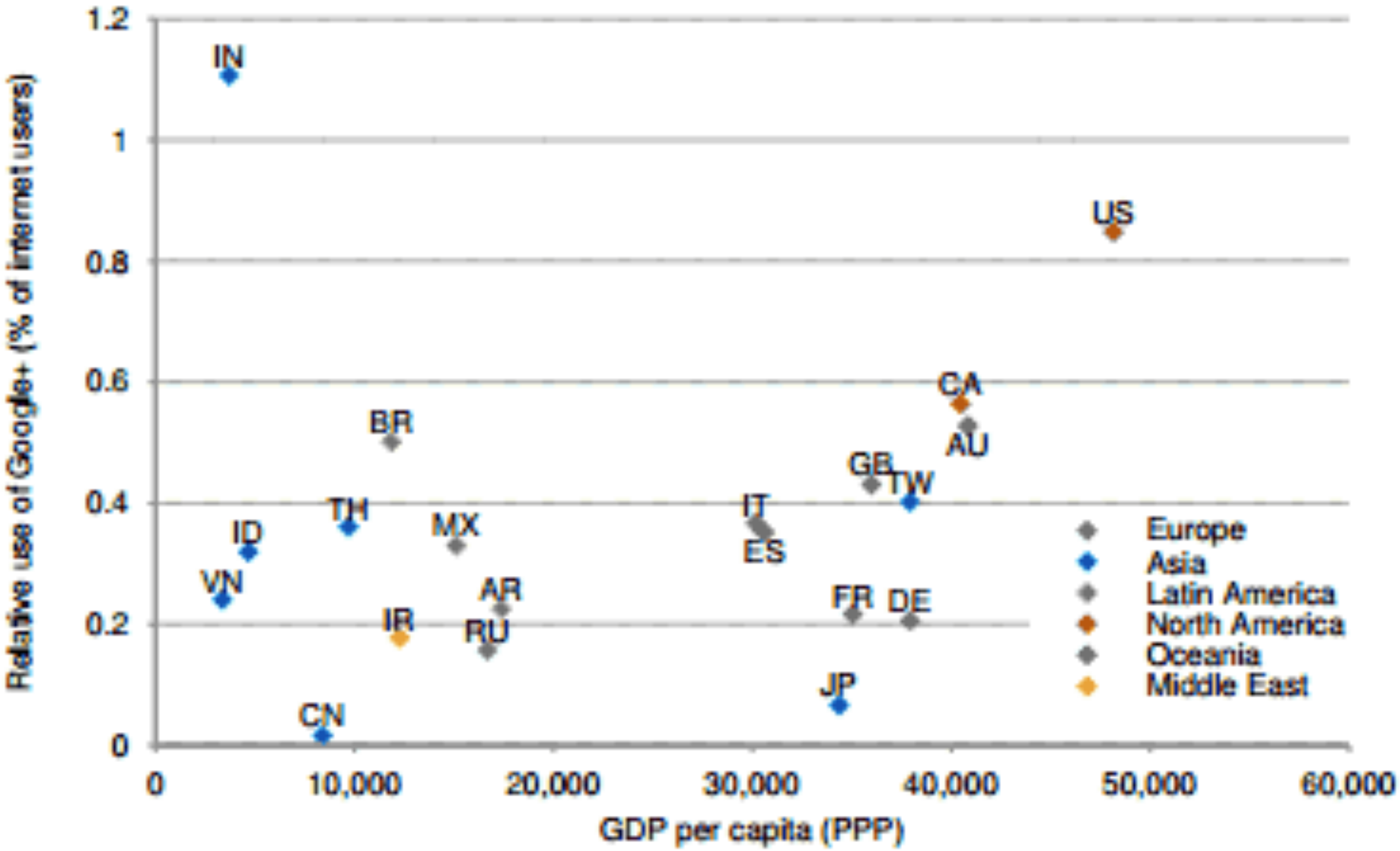


(b) GDP Per Capita and Internet Penetration

country shares among Google+ users

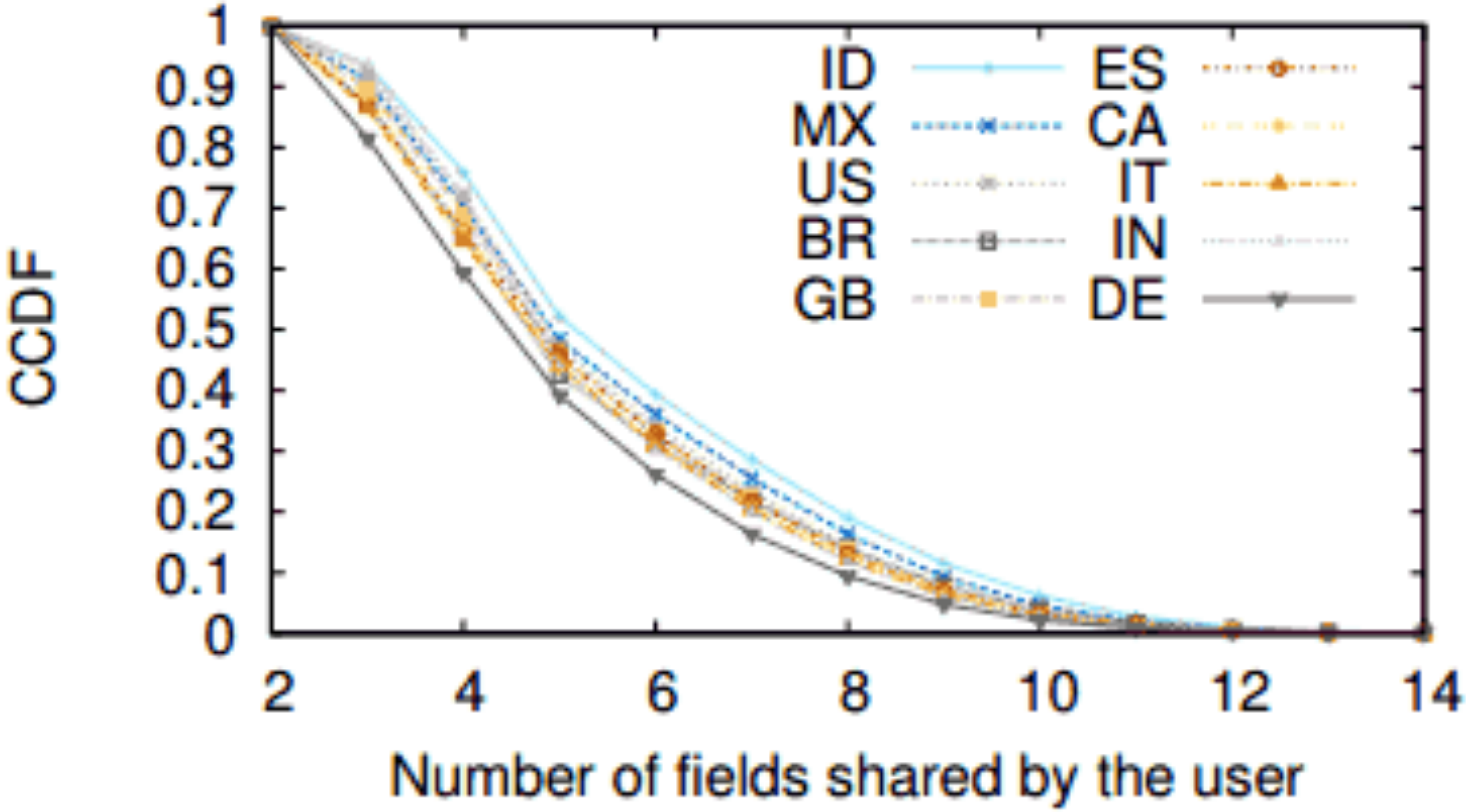


Google+ adoption

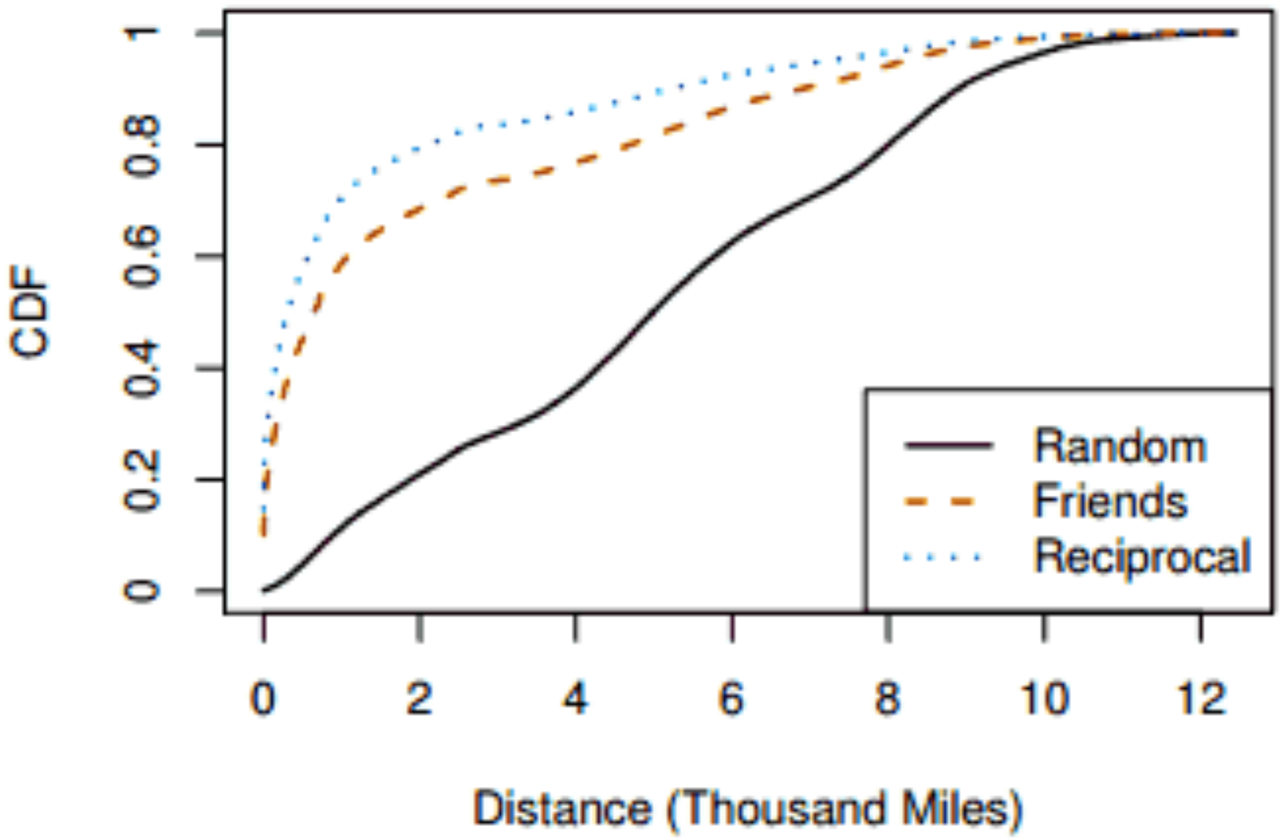


(a) GDP Per Capita and Use of Google+

information sharing by country

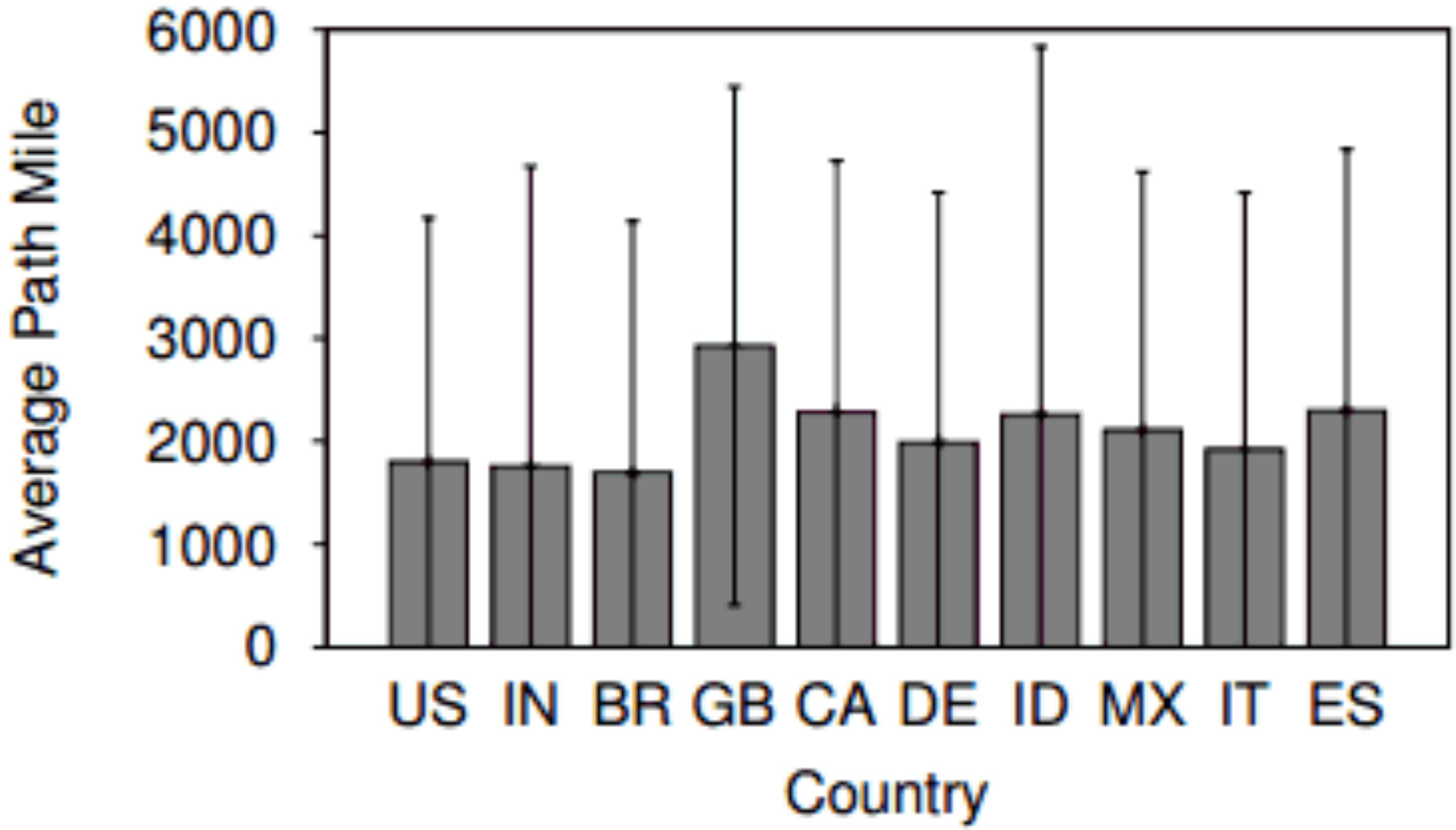


path mile distribution of Google+ users



(a) Path Mile distribution of Google+ users

average path mile by country



(b) Average path mile with standard deviation

link distribution across countries

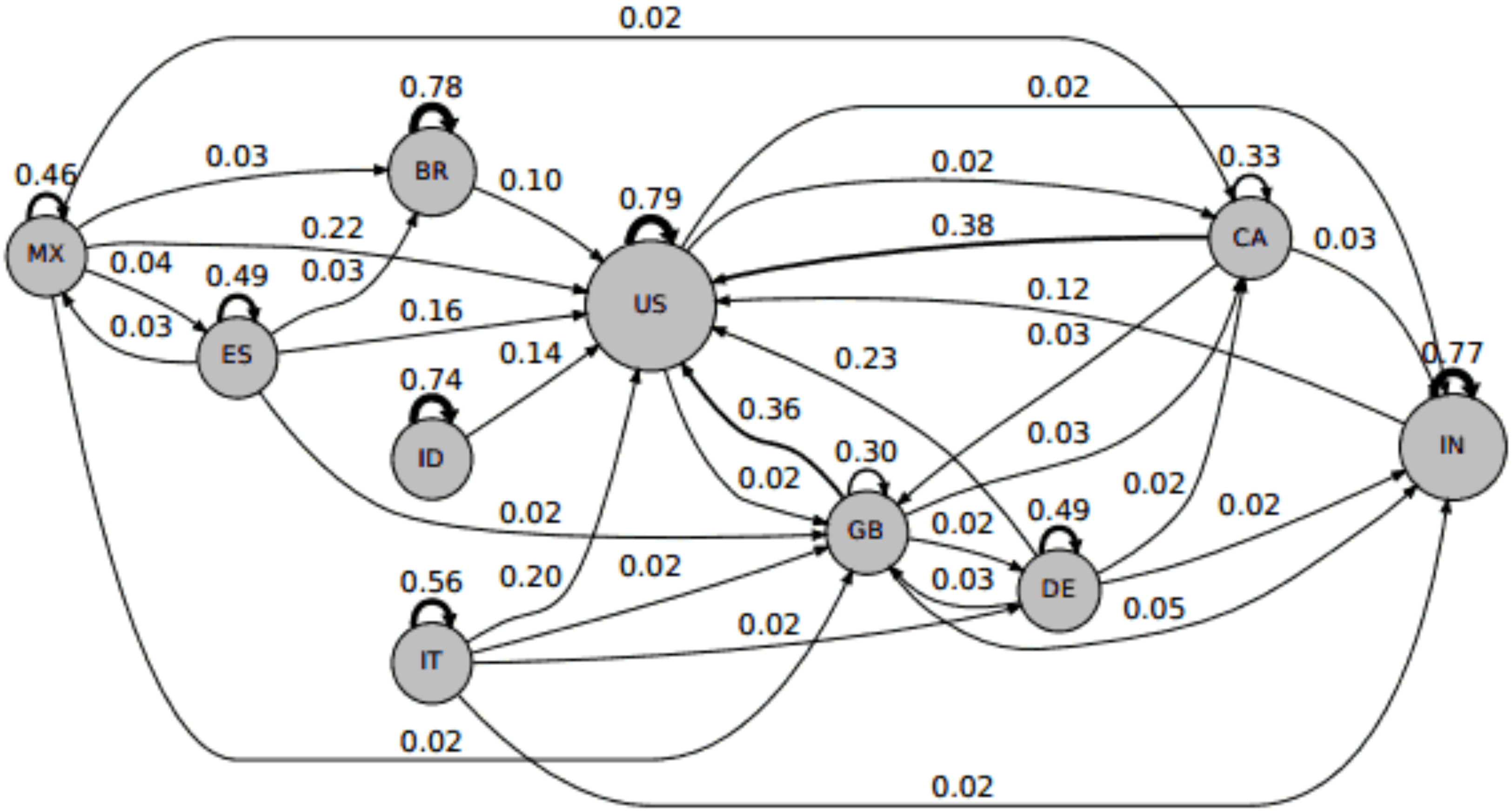


Figure 10: Link distribution across the top countries

conclusions (from paper)

- shorter path lengths, but probably due to size and novelty (?)
- penetration info useful for marketing and for guiding expansion
- significantly different privacy behavior across countries
- significantly different inward-looking/ outward-looking behaviors

observations

- longitudinal observations would make this significantly more useful
- structure of network depends on complex cultural, social, linguistic, etc. factors
- analysis by machine learning, e.g. predict probability of a link between two users based on distance, country, languages, profession, etc.

Evolution of Social Attribute Networks

Gong et. al, IMC 2012

main goals

Creation of a Google+ dataset

Capture both state and evolution over time

Evaluation and evolution of metrics

Creation of a generative model

social attribute networks

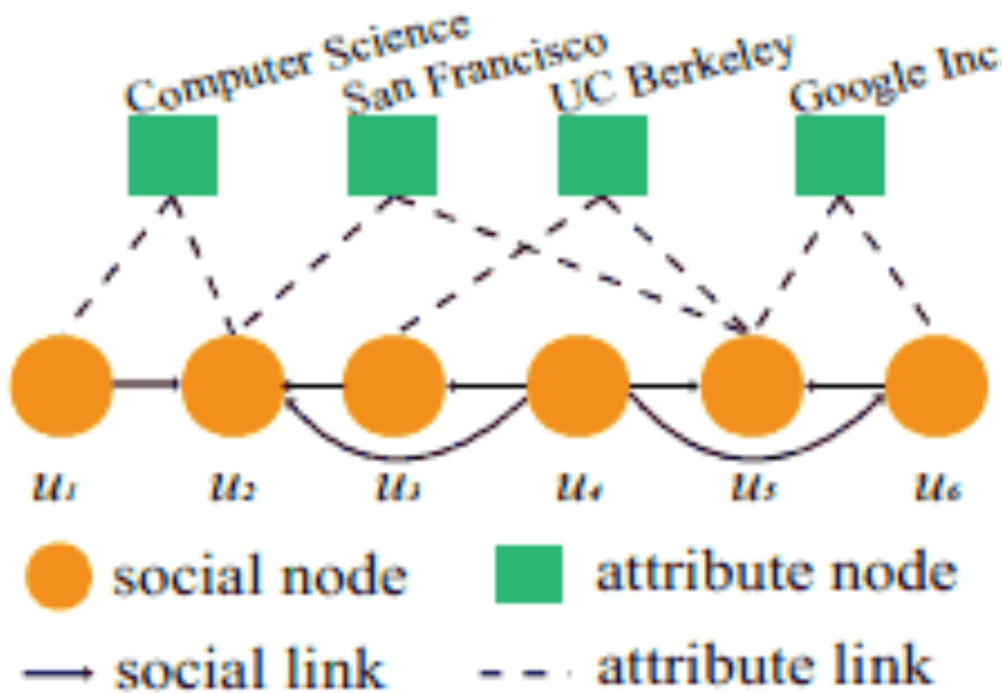
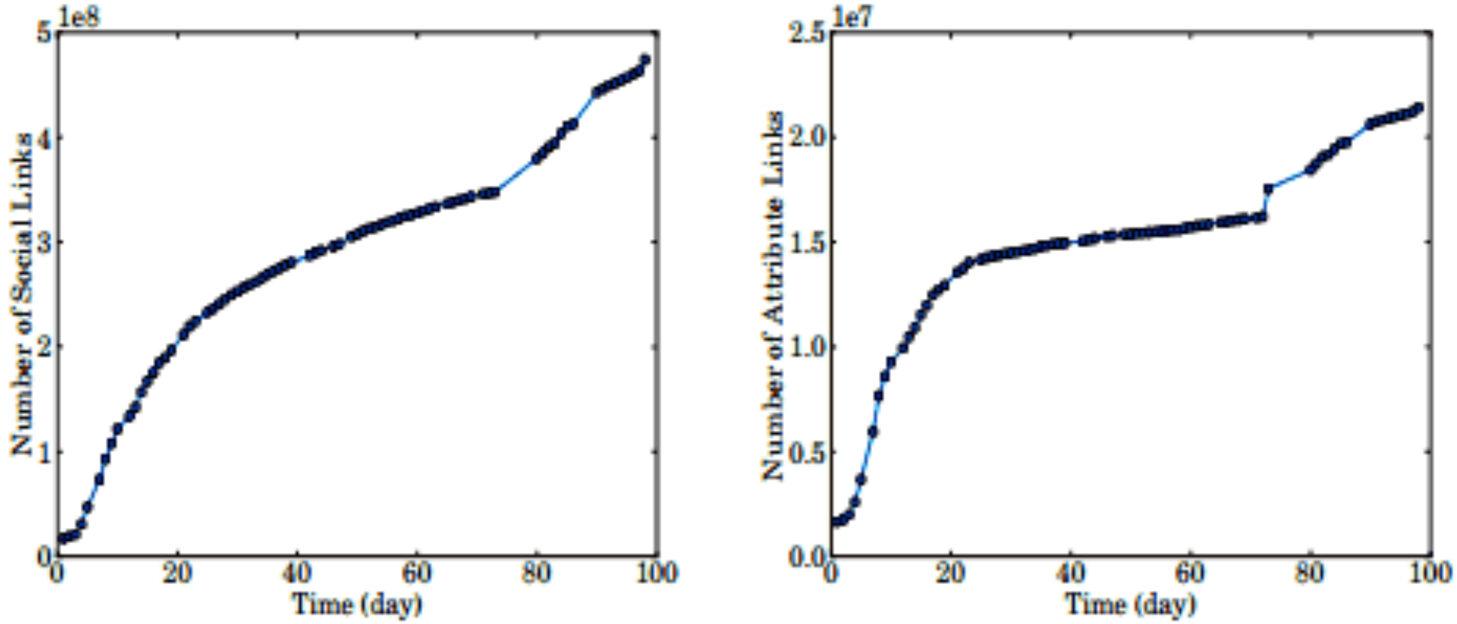


Figure 1: *Illustration of a SAN with six social nodes and four attribute nodes. Note that the social links between users are directed whereas the attribute-user links are undirected.*

network structure

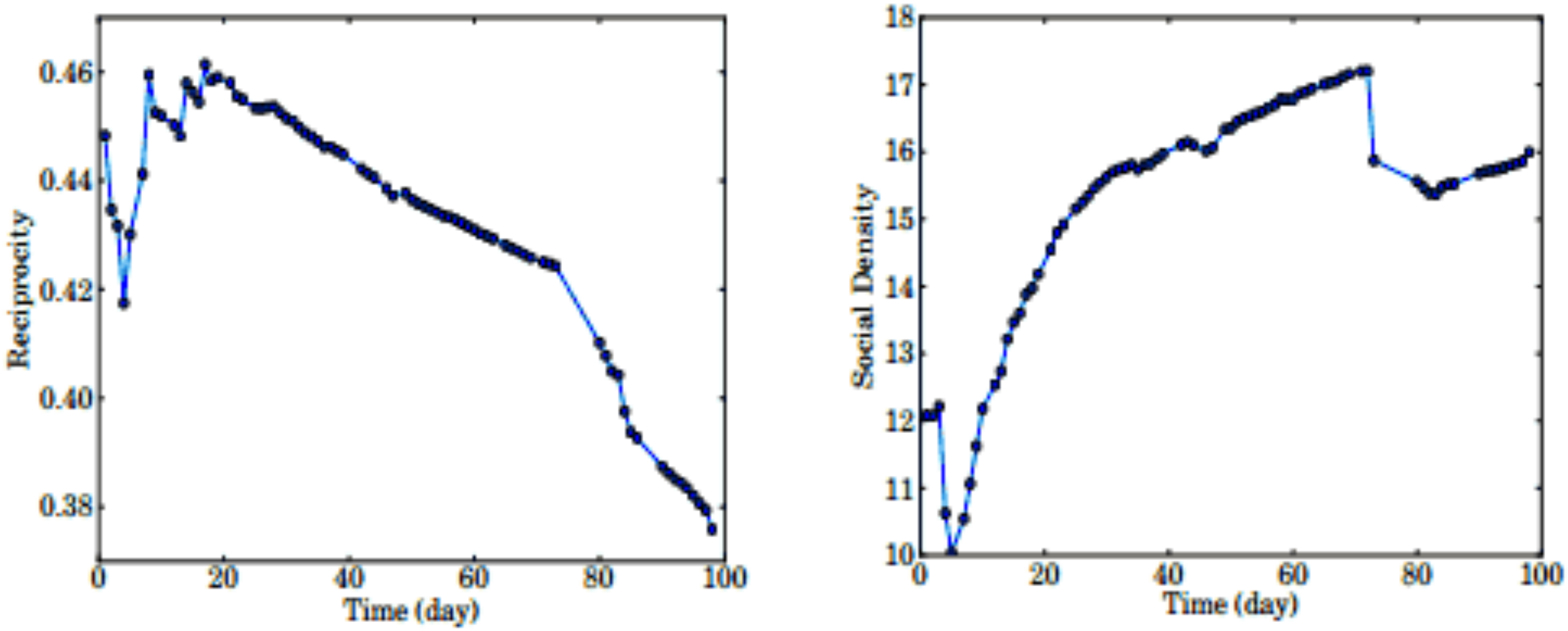
growth in # links



(a) Social links (b) Attribute links

Figure 3: *Growth in the number of social and attribute links in the Google+ dataset*

evolution of reciprocity and social density



(a) Reciprocity (b) Social density

Social density = ratio of links to nodes. Increasing in citation and affiliation networks, Facebook. Fluctuates on Flickr, fairly constant on E-mail.

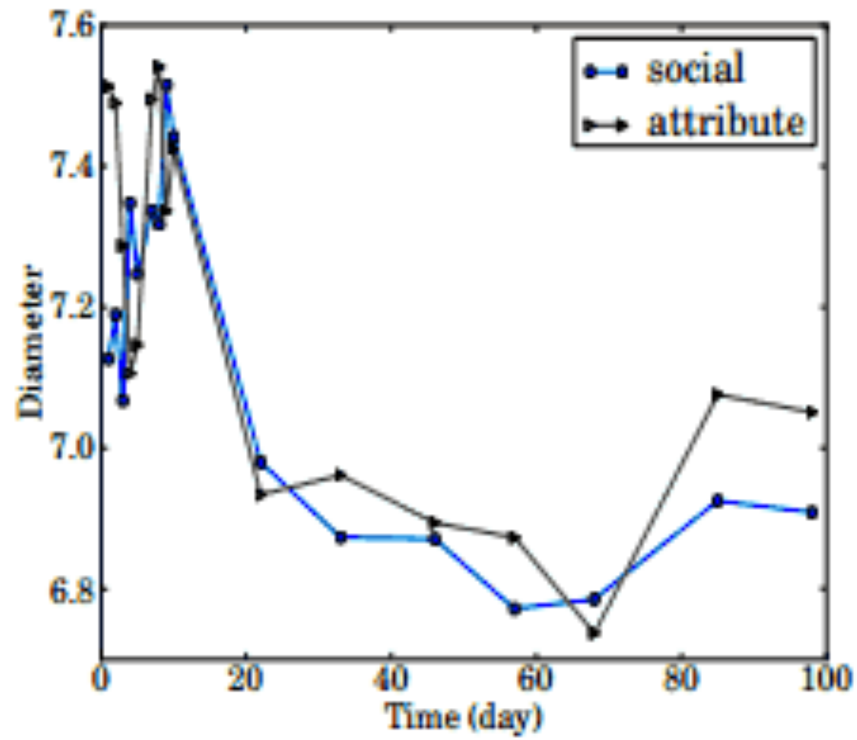
Phases: (1) many people join, but not many friends yet, (2) like social networks, (3) public release of Google+ causes drop, then rises as before.

evolution of diameter

same three phases

not huge differences

for most networks, shrinks over time

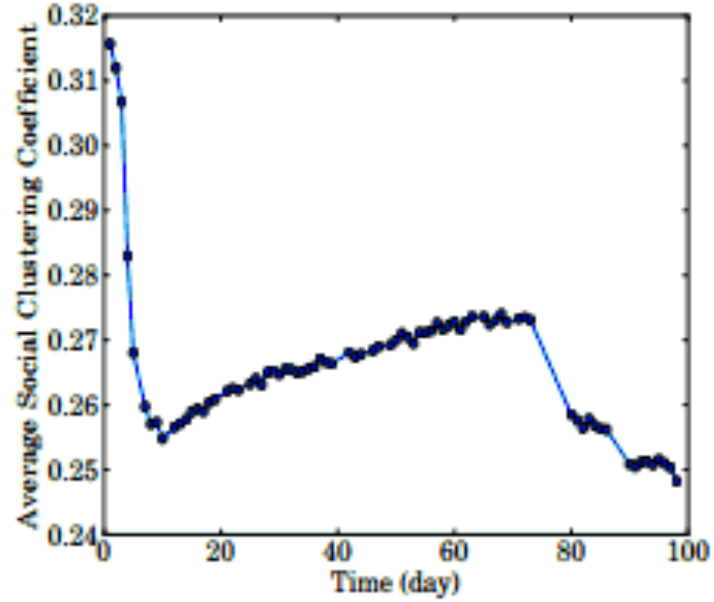


(c) Diameter

clustering

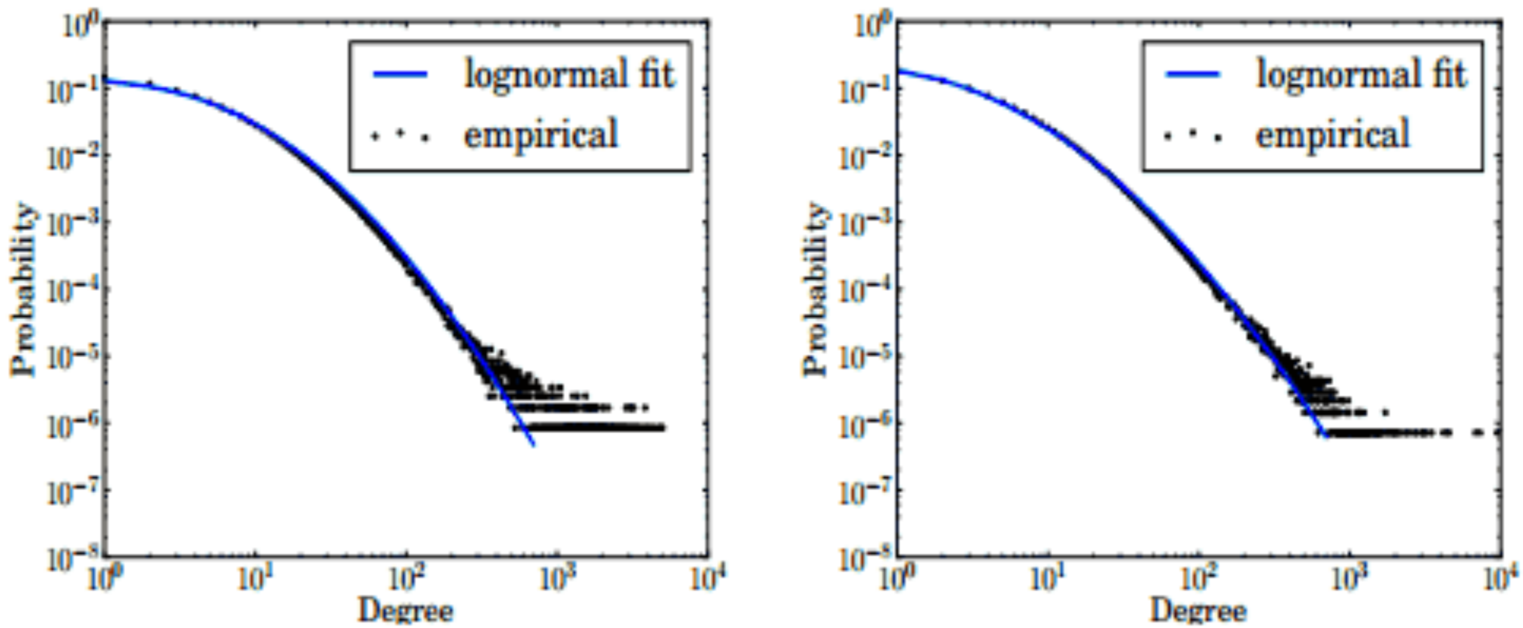
constant for E-mail,
unknown for others

slow increase in phase 2,
suggesting community
formation



(d) Social clustering coefficient

distribution of in/outdegrees

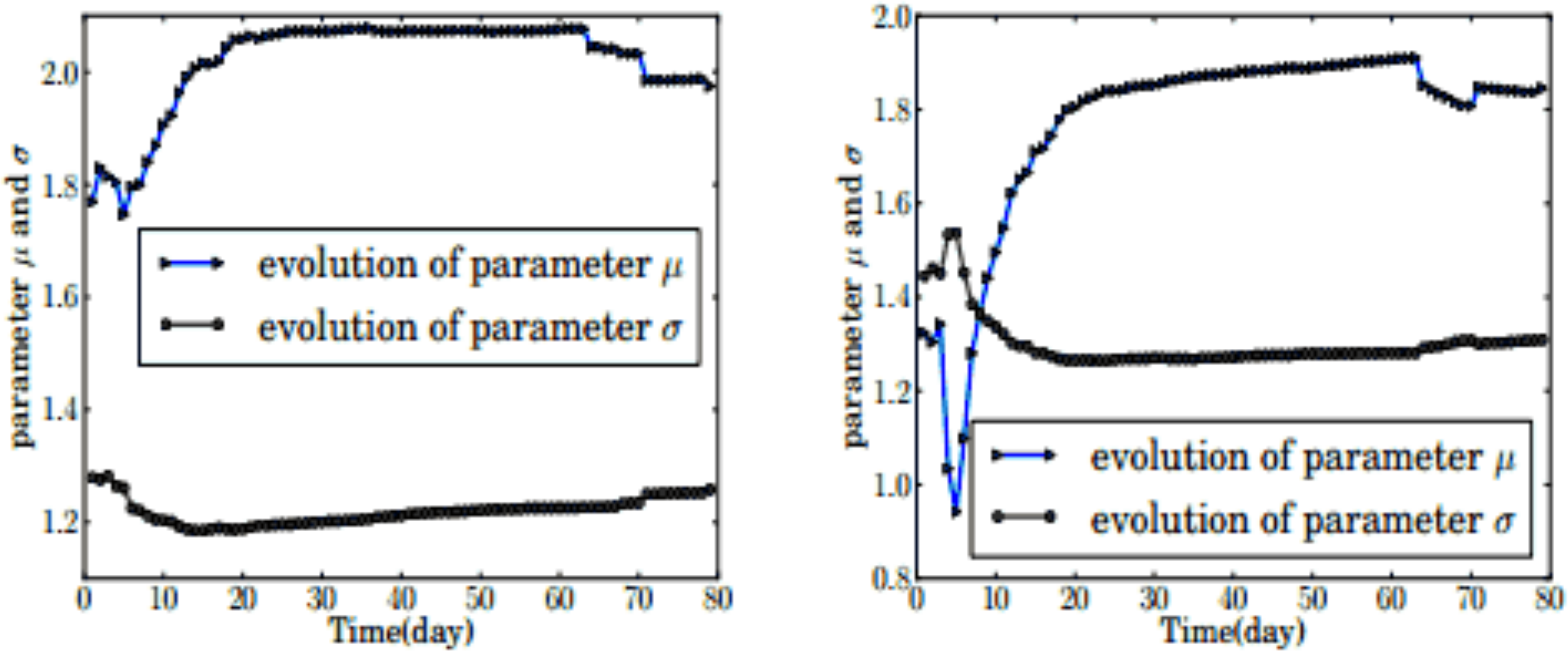


(a) Outdegree (b) Indegree

Figure 5: *Indegree and outdegree distributions for the social nodes in the Google+ SAN along with their best-fit curves. We observe that both are best modeled by a discrete lognormal distribution unlike many networks that suggest power-law distributions.*

where have we seen a log normal distribution before?

evolution of degree distributions



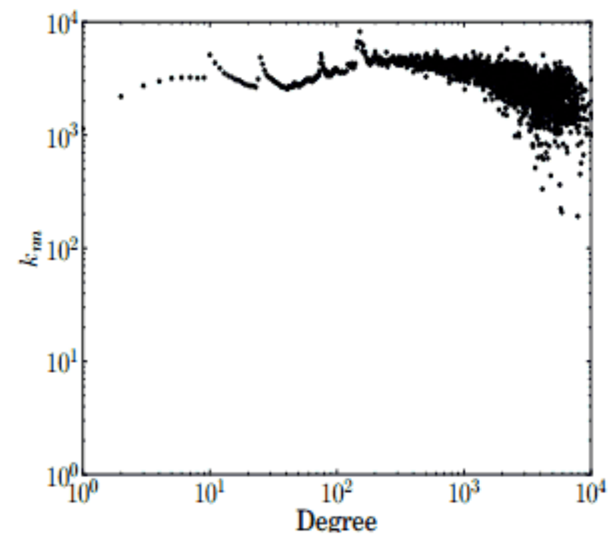
(a) Outdegree (b) Indegree

Figure 6: *Evolution of the lognormal parameters for the indegree and outdegree distributions.*

joint degree distribution

average in-degree of nodes attaching to nodes with a given out-degree

increasing = high-degree tends to connect to high-degree



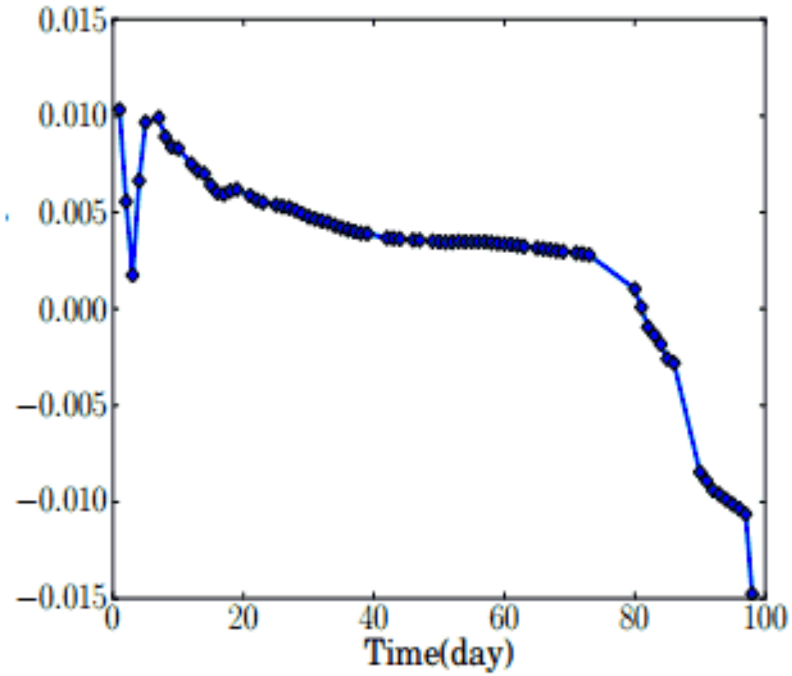
(a) k_{nn} metric

assortativity

correlation coefficient of JDD

typically positive for social networks, negative for pub/sub

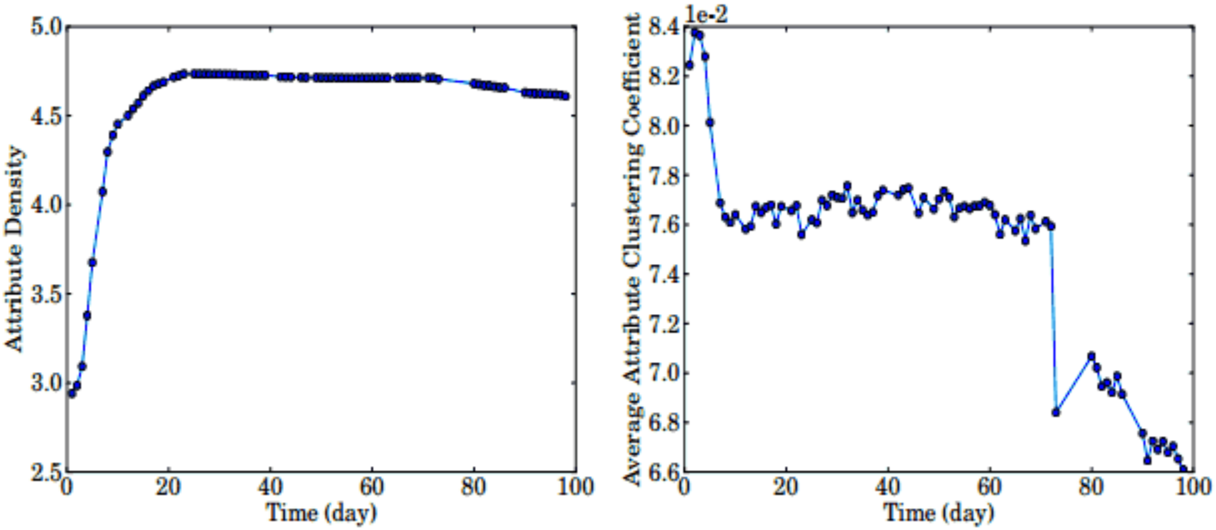
near 0 for Google+ = hybrid network?



(b) Assortativity

attribute structure

attribute density & clustering



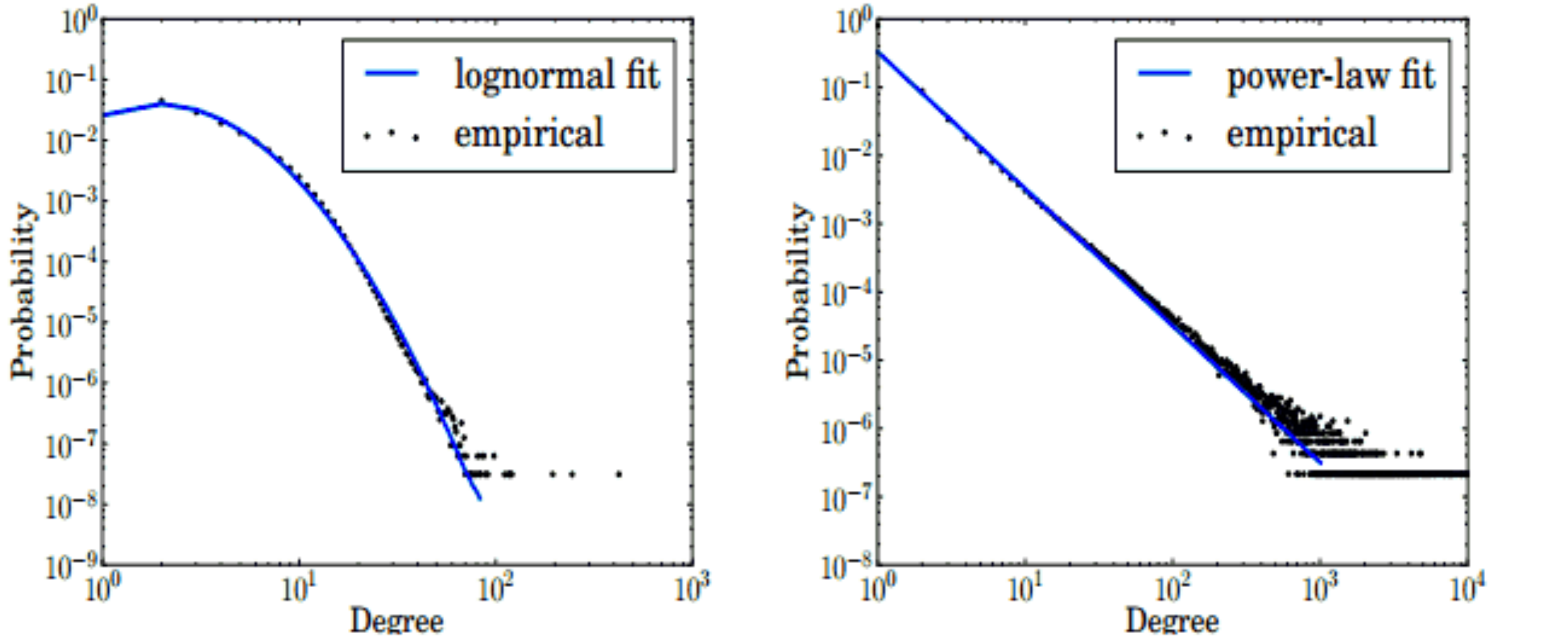
(a) Attribute density (b) Clustering coefficient

Attribute density: |Ea|/|Va|

Clustering density for attributes: “third link” is social link, so propensity for attributes to form communities.

NB: this includes attributes like “city” etc.

distributions of attribute-induced degrees

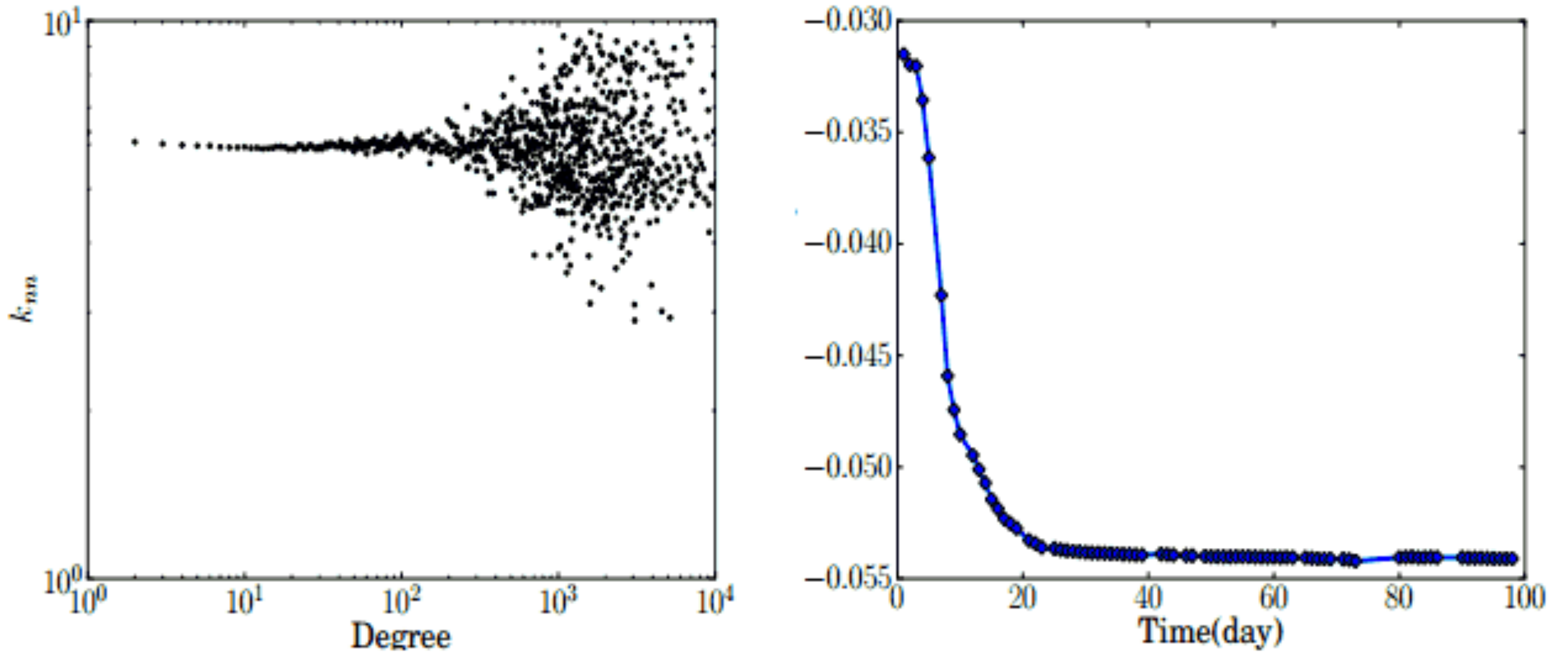


(a) Attribute degree of social nodes

(b) Social degree of attribute nodes

Figure 10: *Distributions of attribute-induced degrees in the Google+ SAN along with their best fits. The attribute degree of social nodes is best modeled by a lognormal whereas the social degree of attribute nodes is best modeled by a power-law distribution.*

assortativity of attributes



(a) Attribute k_{nn} (b) Evolution of assortativity

Figure 12: *(a) Joint degree of attribute nodes: Log-log plot of the social degree versus the average attribute degree of social neighbors of attribute nodes. (b) The evolution of the attribute assortativity coefficient.*

comment… are these well chosen?

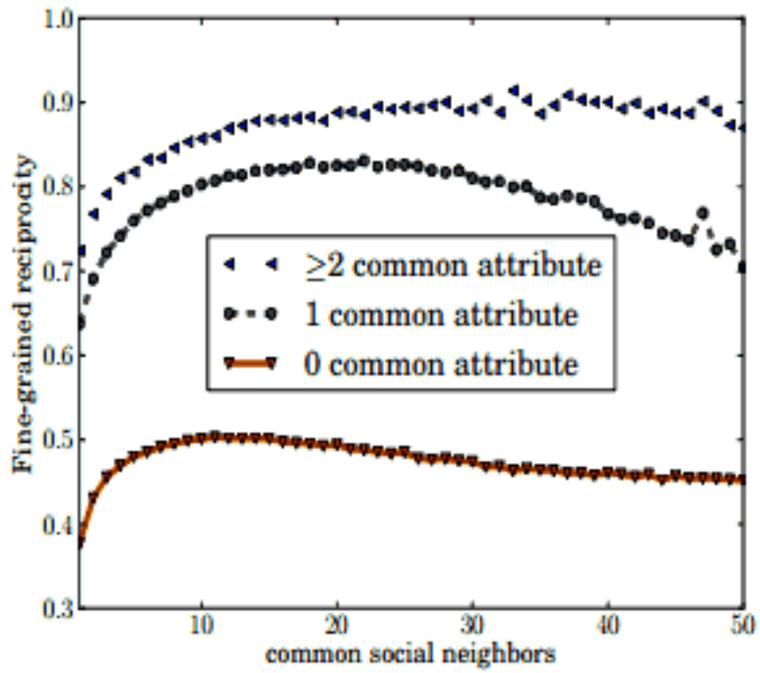
Interesting questions about the relationship between attributes and the social graph.

Why “extend” graph analysis?

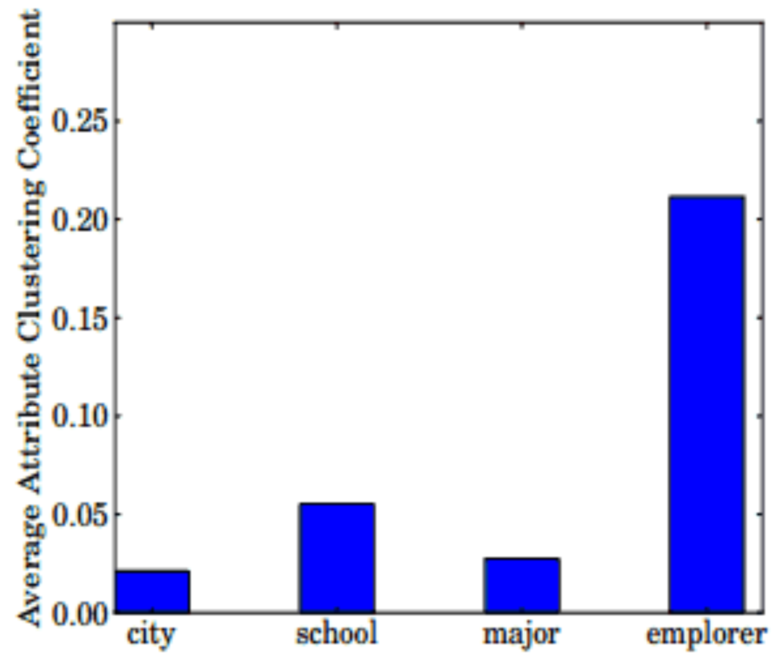
What are some of the things wrong with this approach?

What are better approaches?

social network structure vs attributes



(a) Reciprocity



(b) Clustering coefficients

Figure 13: *Influence of attribute on reciprocity and clustering co-efficients.*

observations on Google+ measurements

- When viewed as a graph structure, the attribute graph has different kinds of distributions and behaviors from the social graph.
- The attribute graph influences the social graph in “interesting” ways.
- Some attributes have stronger influence than others.

(duh!)

generative model

preferential attachment

Build a network by adding nodes one at a time.

Nodes attach to existing nodes with a probability related to their degree. Extension: number of shared attributes *a*.

- *Power Attribute Preferential Attachment (PAPA):*
  $f(u,v) \propto d_i(v)^\alpha (1 + a(u,v)^\beta)$
- *Linear Attribute Preferential Attachment (LAPA):*
  $f(u,v) \propto d_i(v)^\alpha (1 + \beta \cdot a(u,v))$

triangle closing

Friend requests to friends of friends. Do attributes improve triangle closing?

- *Baseline*: Select a social neighbor v within a 2-hop radius uniformly at random.
- *Random-Random (RR)*: Select a social neighbor $w \in \Gamma_s(u)$ uniformly at random, and then select a social neighbor $v \in \Gamma_s(w)$ uniformly at random which is shown to have very good performance in previous work [29].

  14% better than baseline
- *Random-Random-SAN (RR-SAN)*: select a neighbor $w \in \Gamma_s(u) \cup \Gamma_a(u)$ uniformly at random, and then select a social neighbor $v \in \Gamma_s(w)$ uniformly at random.[4]

  36% better than RR

gamma-a: attribute neighbors, gamma-s: social neighbors

generative model

Algorithm 1: Social-Attribute Network Model

T, simulated time steps
Initialization.
for $1 \leq t \leq T$ **do**
    Social node arrival. Sample a set of new social nodes $V_{t,new}$.
    for $v_{new} \in V_{t,new}$ **do**
        Attribute degree sampling. Sample the number of attributes $n_a(v_{new})$ for v_{new} from a lognormal distribution.
        for $1 \leq i \leq n_a(v_{new})$ **do**
            Attribute linking.
        end
        First outgoing linking.
        lifetime sampling.
        sleep time sampling.
    end
    Collect woken social nodes $V_{t,woken}$.
    for $v_{woken} \in V_{t,woken}$ **do**
        Outgoing linking.
        sleep time sampling.
    end
end

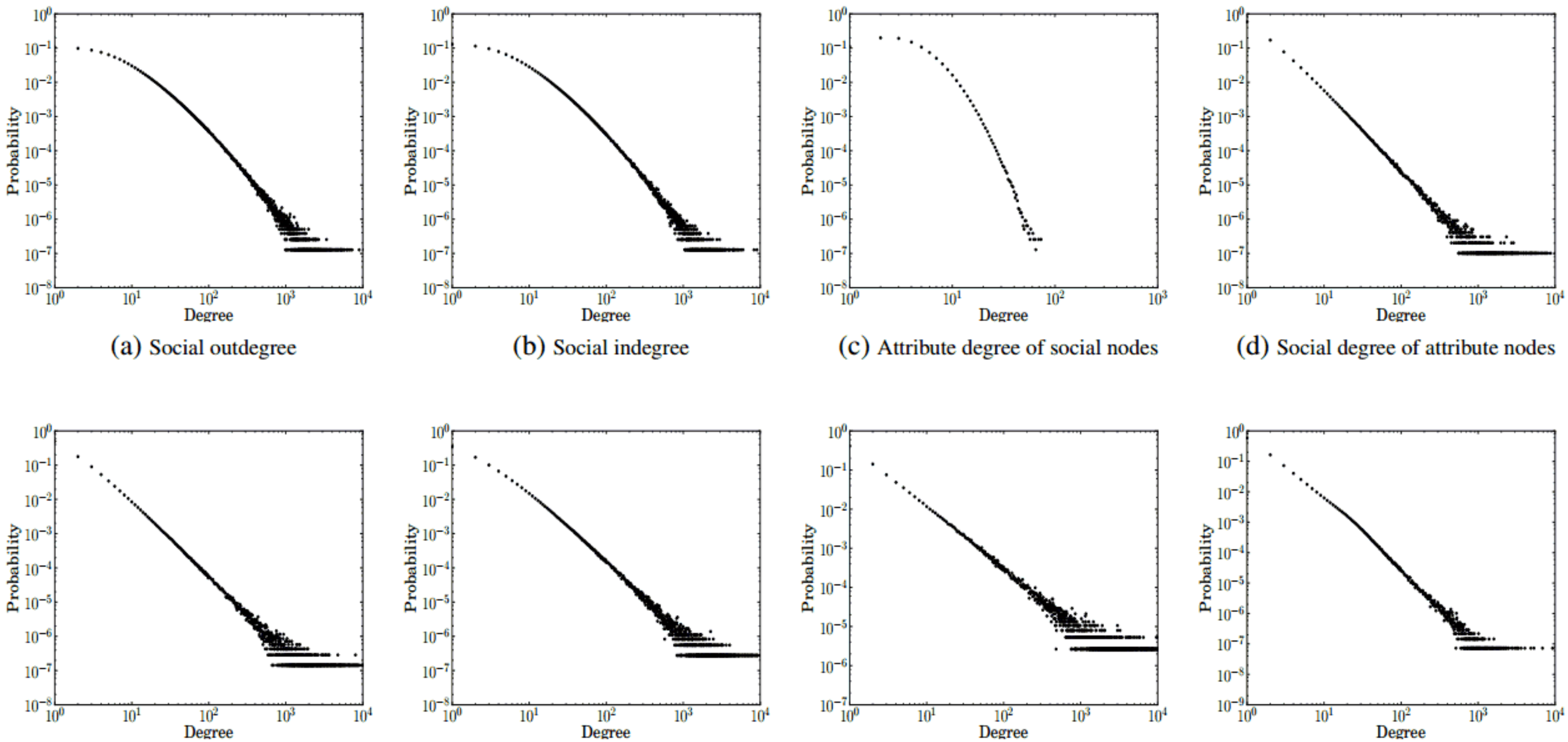
Zheleva model

Algorithm 2 Co-evolution model

Set of nodes $V = \emptyset$
Set of groups $H = \emptyset$
for each time period $t \in T$ **do**
 Set of active nodes at time t, $V_t = \emptyset$
end for
for each time period $t \in T$ **do**
 Node arrival. $V = V \cup V_{t,new}$
 for each new node $v \in V_{t,new}$ **do**
  Lifetime sampling
  First social linking
 end for
 for each node $v \in V_t$ **do**
  Social linking
  Affiliate linking. v determines n_h, the number of groups to join, sampled from an exponential distribution $\lambda' e^{-\lambda' n_h}$ with a mean $\mu' = \frac{1}{\lambda'} = \rho.\text{degree(v)}^{\gamma}$.
  for $i = 1 : n_h$ **do**
   if $rand() < \tau$ **then**
    Group creation. v creates group h, and forms edge $e_a(v, h, t)$. $H = H \cup \{h_i\}$.
   else
    Group joining. v forms edge $e_a(v, h, t)$. Group h is picked through a friend with probability p_v; otherwise, or if no friends' groups are available, it joins a random group with prob. proportional to the size of h.
   end if
  end for
 end for
 for each node $v \in V_t \cup V_{t,new}$ **do**
  Sleep time sampling
 end for
end for

KDD'09, June 28–July 1, 2009, Paris, France.

statistical comparison Gong et al. vs Zhel



(a) Social outdegree
(b) Social indegree
(c) Attribute degree of social nodes
(d) Social degree of attribute nodes
(e) Social outdegree
(f) Social indegree
(g) Attribute degree of social nodes
(h) Social degree of attribute nodes

Figure 16: *Degree distributions of synthetically generated SAN using our model in (a)-(d) vs.* Zhel *shown in (e)-(h).*

statistical comparison Gong et al. vs Zhel

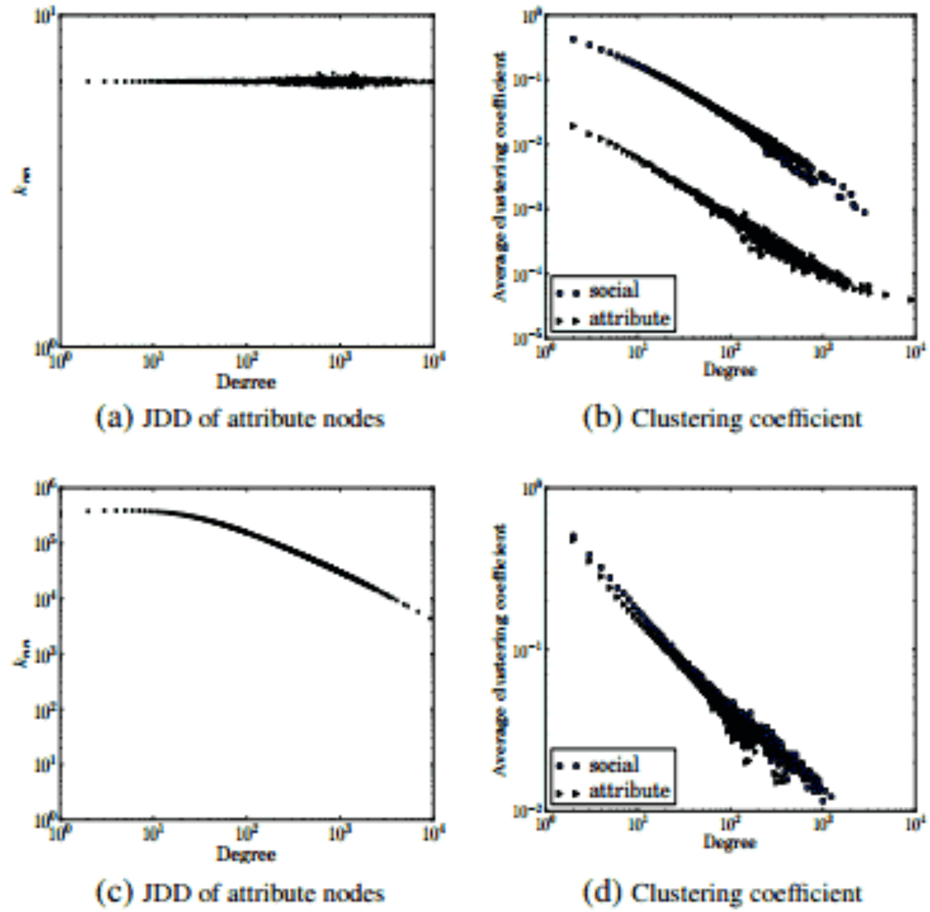


Figure 17: *Joint degree and clustering coefficient distributions of our model (a)–(b) vs.* Zhel *in (c)–(d).*

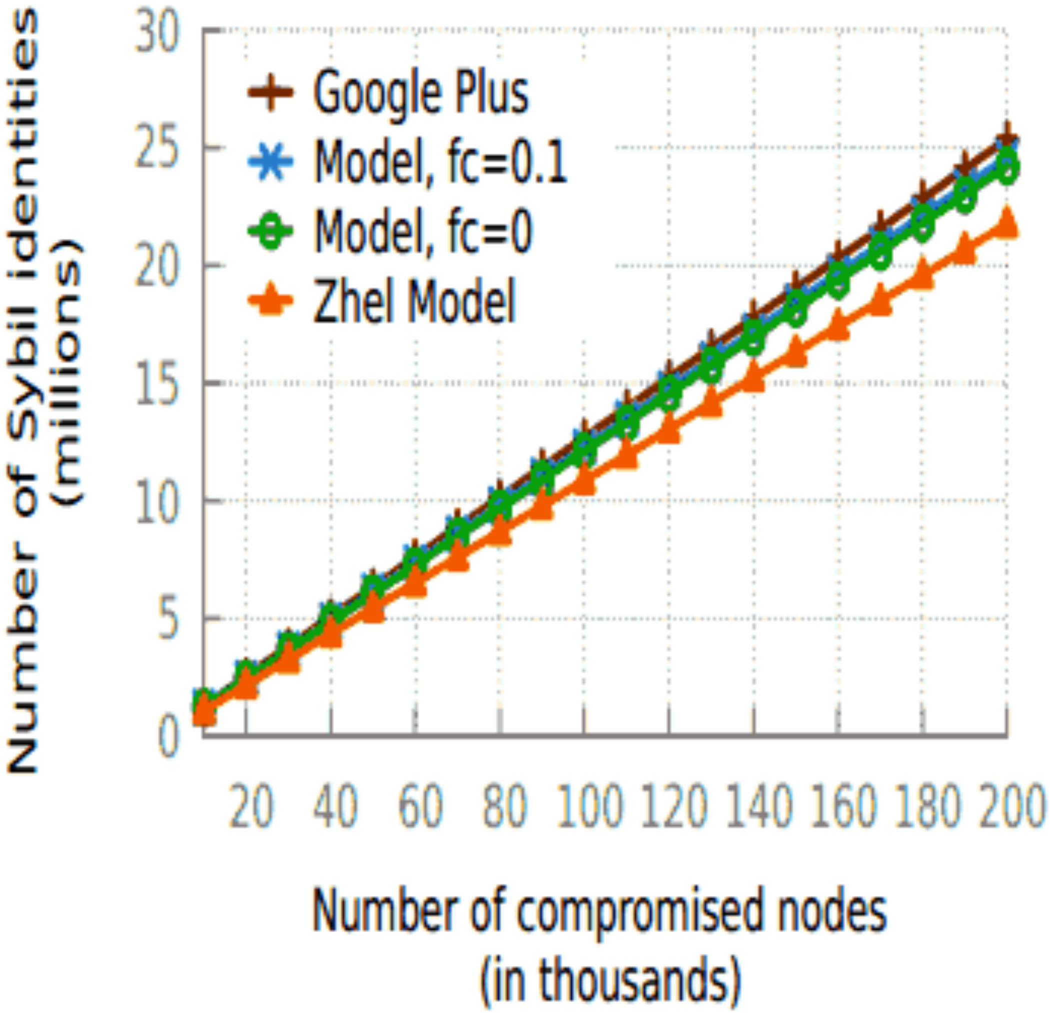
application validation: Sybil Defense

Sybil attack (Wikipedia).

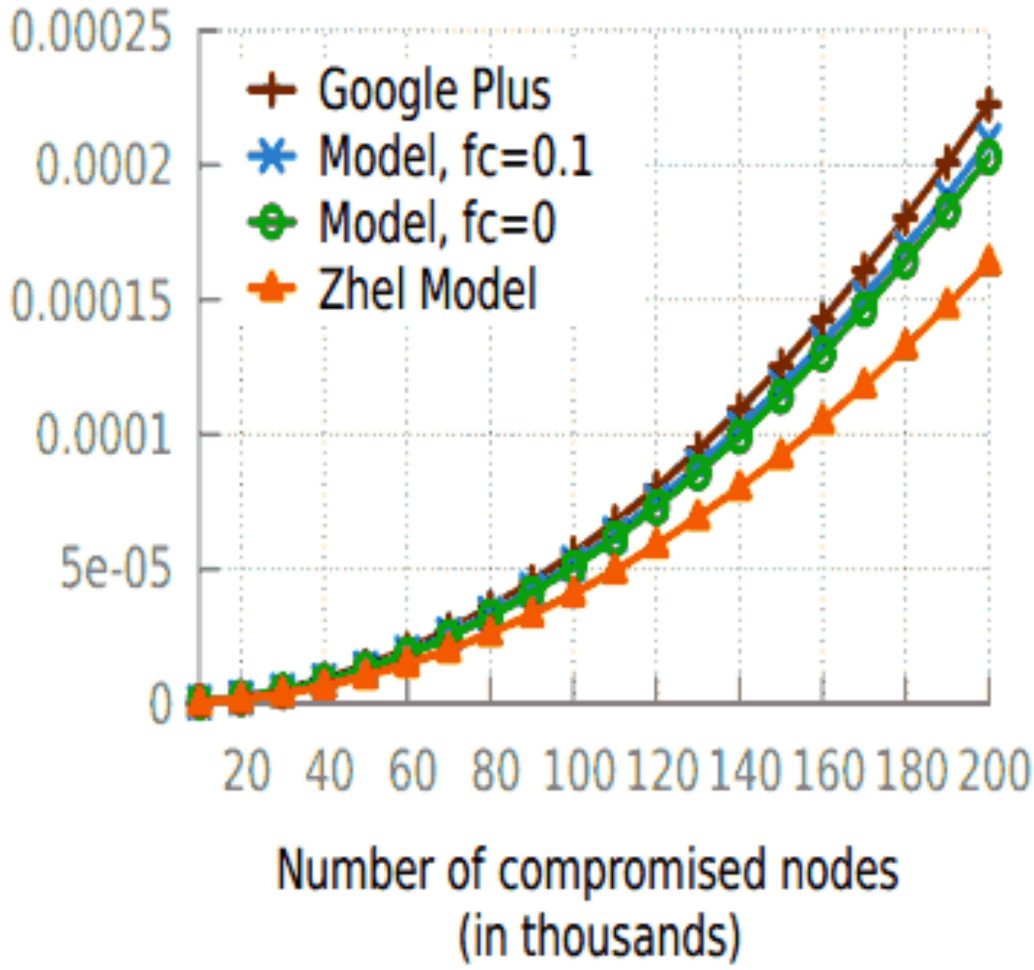
In a Sybil attack the attacker subverts the reputation system of a peer-to-peer network by creating a large number of pseudonymous identities, using them to gain a disproportionately large influence. (AKA “sockpuppets”)

Simple Sybil defense: limit node connectivity.

application performance prediction with different models



application validation: subverting onion routing on social graph



generative model summary

- model matches statistics of Google+ better than Zhel model
- model performs better for two tested applications

generative model - purpose

We build generative models for several reasons:

- validate our understanding of the mechanisms building up the network
- create data for predicting application performance
- model the effects of interventions