Information Sharing and Designed Social Systems

Jon Kleinberg

Cornell University

Including joint work with Lada Adamic, Lars Backstrom, Cristian Danescu-Niculescu-Mizil, Justin Cheng, Alex Dow, Lillian Lee, and Johan Ugander.

Managing Social Information

Two tensions in the on-line world:

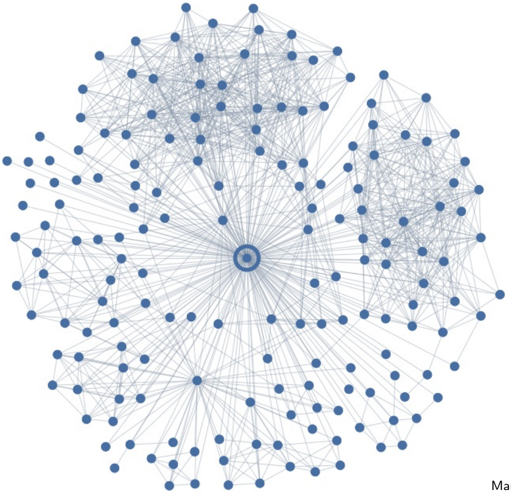
- Library vs. crowd.
- Organic vs. designed.

Designed Social Systems

Algorithmic management of socially shared information:
Facebook as a designed social system

- Which features should be deployed?
  [Ugander-Karrer-Backstrom-Kleinberg 2013]
- Which discussions will be most active?
  [Backstrom-Kleinberg-Lee-DanescuNiculescuMizil 2013]
- Which memes will receive the most reshares?
  [Cheng-Adamic-Dow-Kleinberg-Leskovec 2014]
- Which links should be emphasized?
  [Backstrom-Kleinberg 2014]

Network Neighborhoods

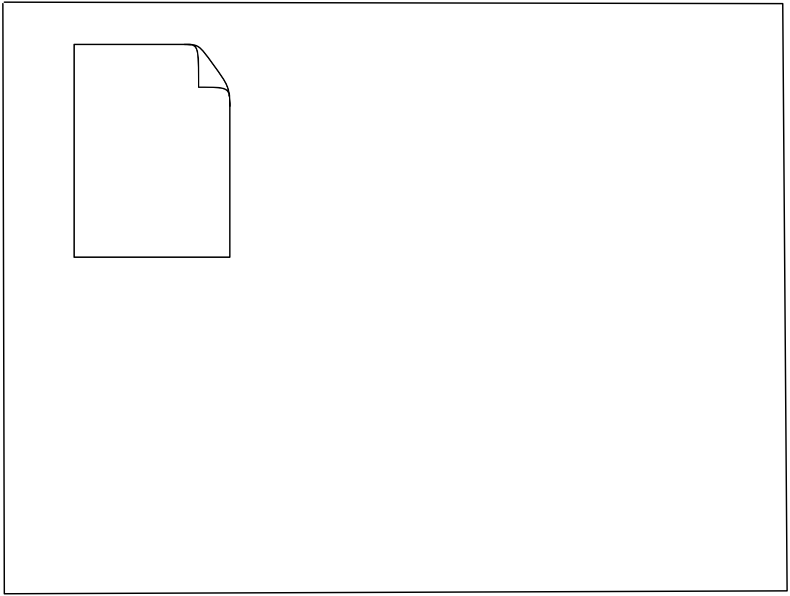


Marlow-Byron-Lento-Rosenn 2009

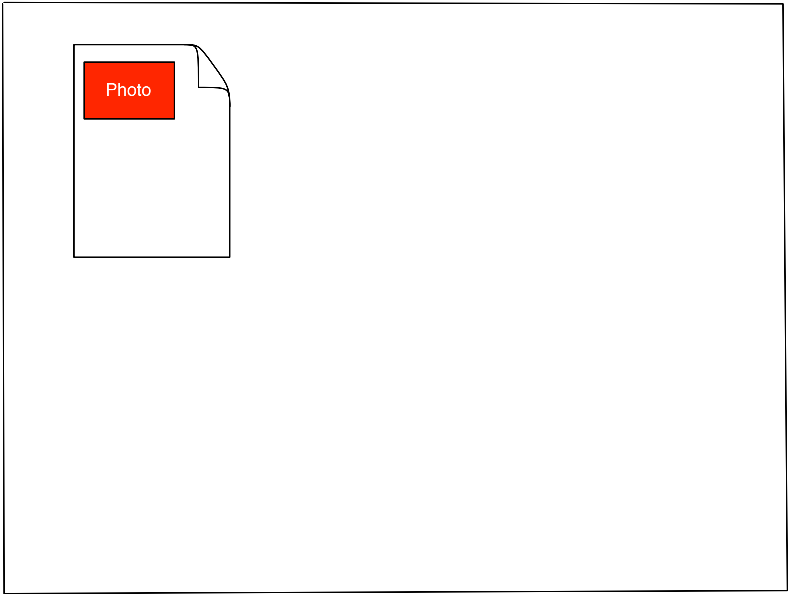
One person's network neighborhood:

- The "input" for their experience in a social-networking system (cf. [Ugander et al 2012, 2013])

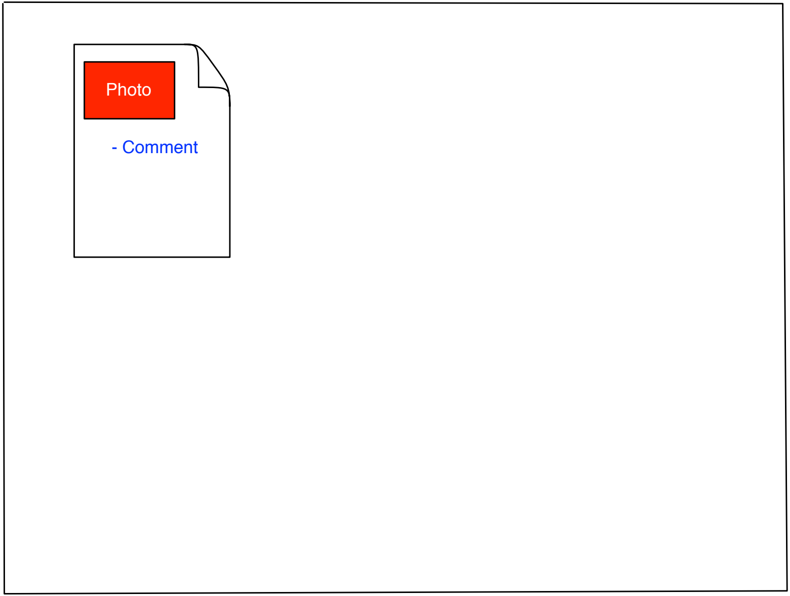
Socially Shared Information



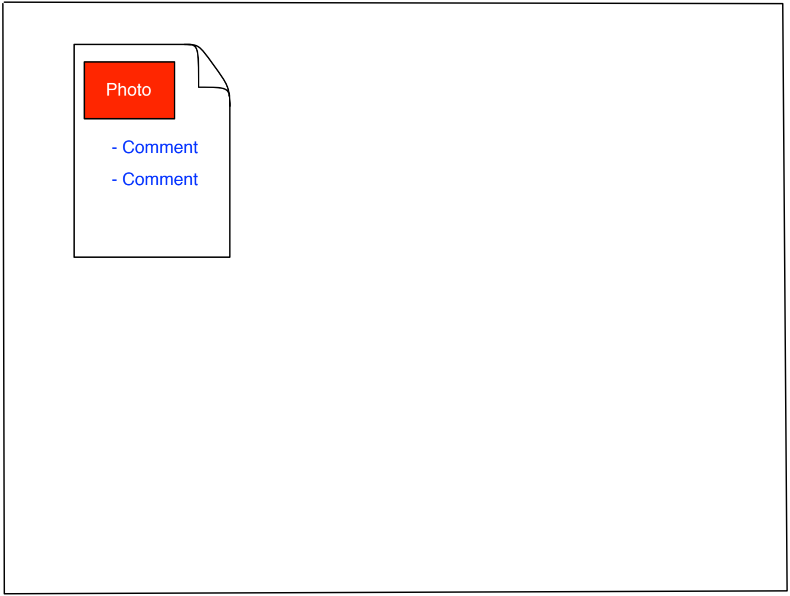
Socially Shared Information



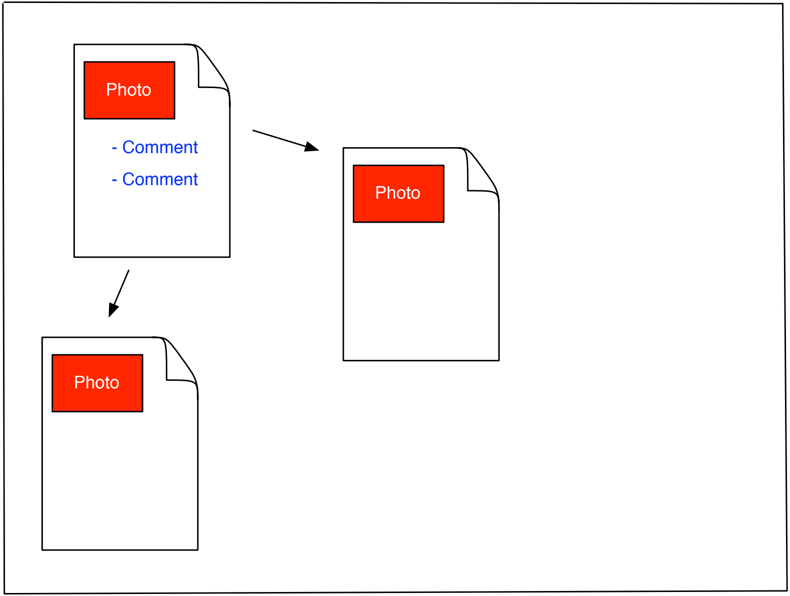
Socially Shared Information



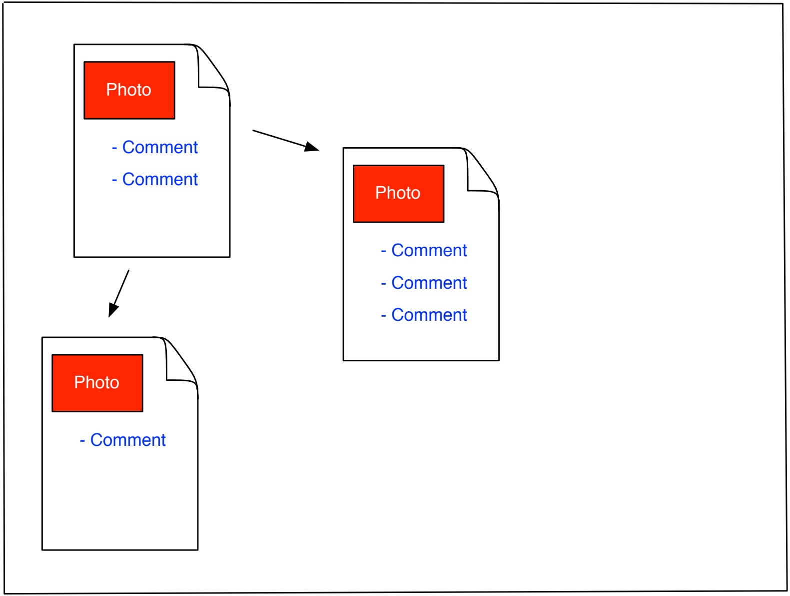
Socially Shared Information



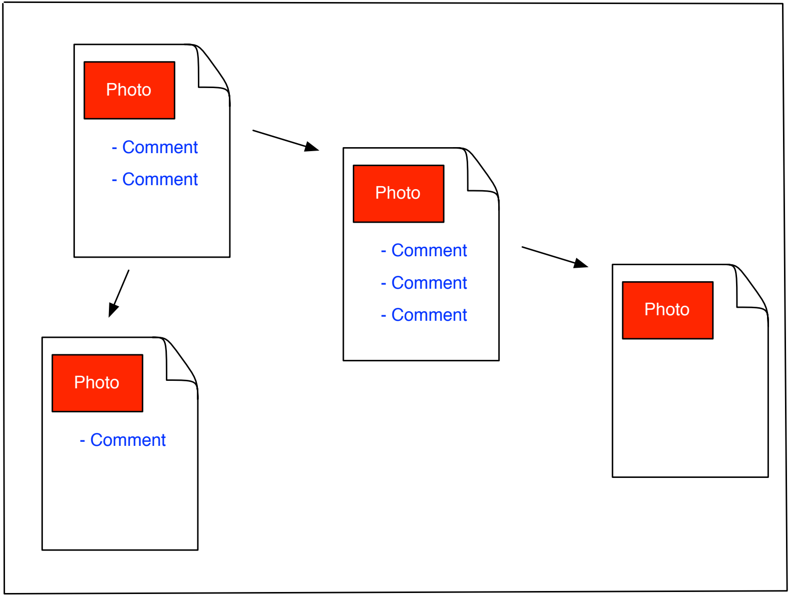
Socially Shared Information



Socially Shared Information



Socially Shared Information



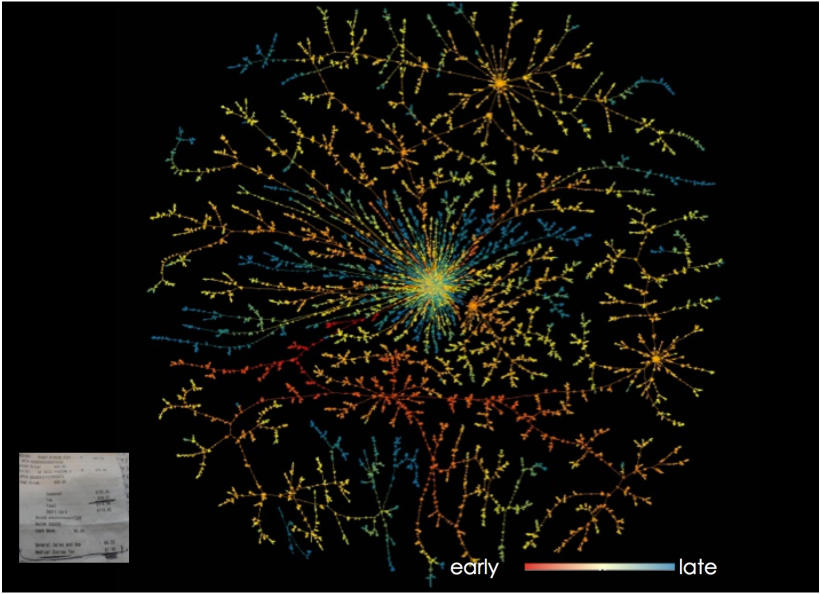
Socially Shared Information

Socially Shared Information

- Comment

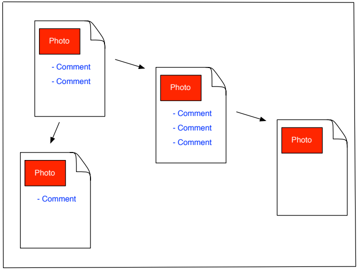
- Comment

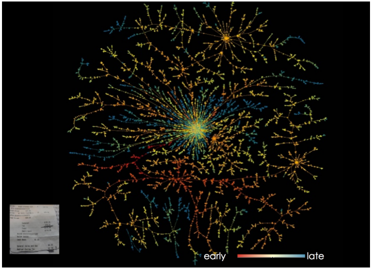
- Comment



Adamic-Cheng-Eckles-Friggeri 2014

Basic Prediction Task





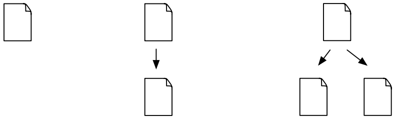
Given the trajectory up to a certain point, predict eventual size.

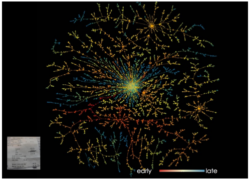
- Can do this for comment threads [Backstrom et al 2013] and reshare cascades [Cheng et al 2014].
- Heuristic for quickly finding most popular content.

A strong challenge: are cascades inherently unpredictable?

- [Salganik-Dodds-Watts 06, Goel et al 12]

Basic Prediction Task





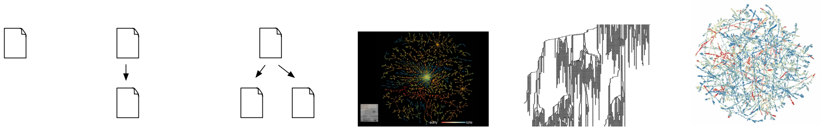
Large cascades are rare but important [Adamic-Dow-Friggeri 2013].

- Most photos are never reshared; almost all cascades are very small.
- But half of all reshares occur in cascades of size > 500.

Challenge for defining a prediction task.

- Pure size estimation has a pathological answer ($= 1$).
- Creating a balanced dataset leads to an artificial task.

Basic Prediction Task



Large cascades are rare but important [Adamic-Dow-Friggeri 2013].

- Most photos are never reshared; almost all cascades are very small.
- But half of all reshares occur in cascades of size > 500.

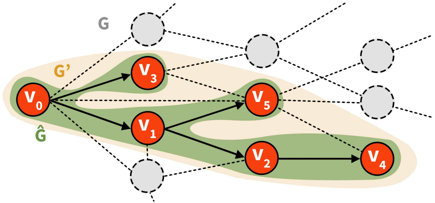
Challenge for defining a prediction task.

- Pure size estimation has a pathological answer $(= 1)$.
- Creating a balanced dataset leads to an artificial task.

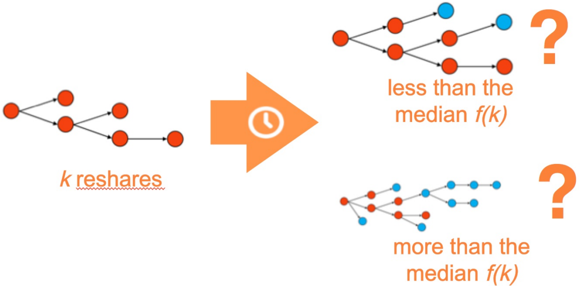
Defining a Prediction Task

Cascade growth prediction.

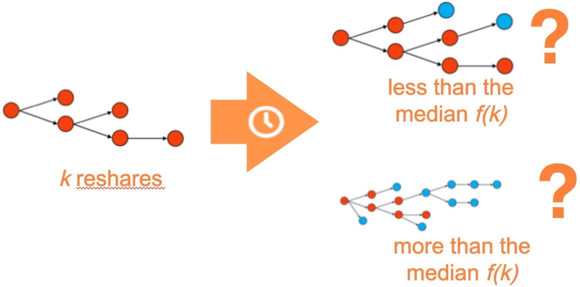
- Let $f(k)$ be median size of cascade conditional on reaching size k.
- Observation on reshare cascades: $f(k) \approx 2k$ for all k.



Given a cascade up to a certain point in time, of size k, predict whether it will reach size $f(k)$.



Cascade Growth Prediction

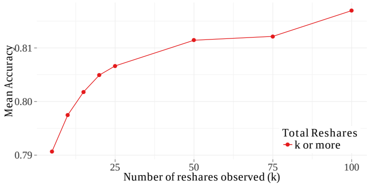


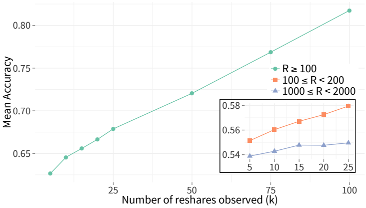
Categories of features:

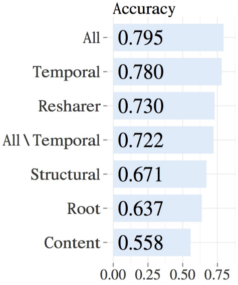
- Content (text overlaid on photo)
- Root (degree, activity level)
- Resharers (degrees, activity levels)
- Structural (initial tree depth, escapes root neighborhood)
- Temporal (time to reach first k, acceleration)

Does the problem get harder or easier with increasing k?

Cascade Growth Prediction



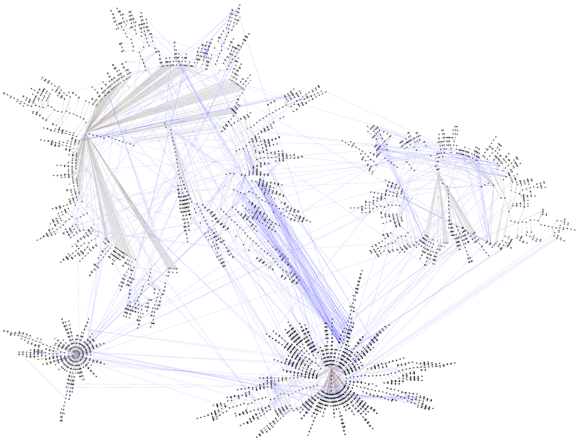




Some general observations:

- Accuracy increases with k.
- Temporal features very powerful.
- High resharer depth predicts larger growth.
- Features of content and original poster get less important with increasing k.

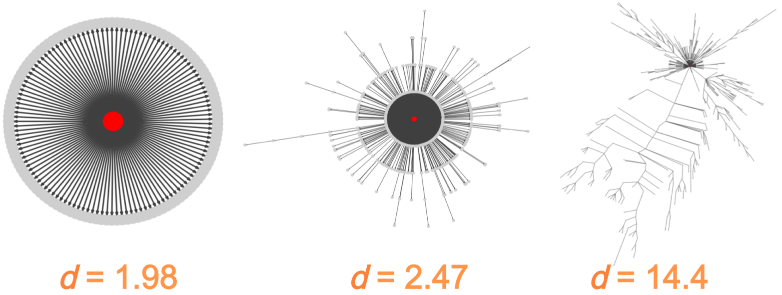
Multiple Cascades



Control for content: Pick 10 random copies of the same photo.

- Given prefix of each, which will produced the largest cascade?
- Random baseline is 10% accuracy.
- Prediction model achieves 49.7%

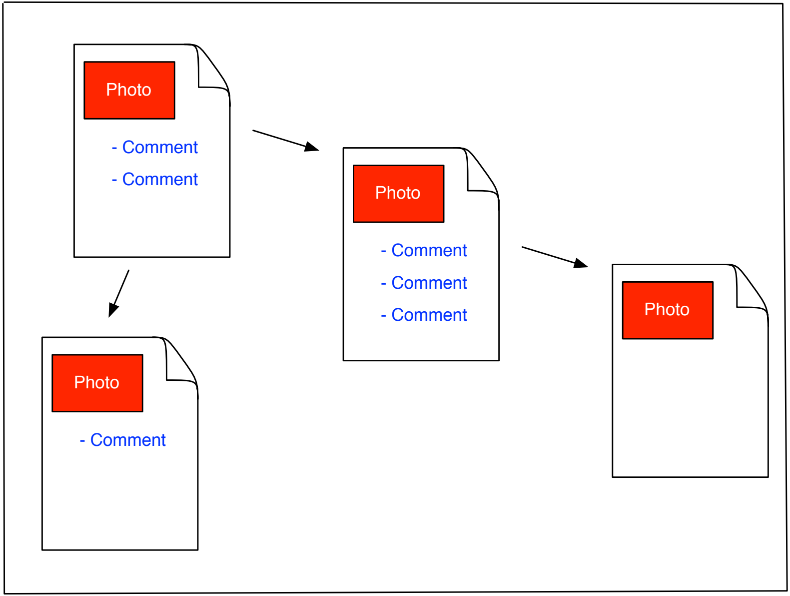
Predicting Structure



Can we predict structural properties of the eventual cascade?

- Wiener index: average distance between nodes in the tree [Anderson-Goel-Hofman-Watts 2014]
- Predict whether this will be above or below median.
- Accuracy of 72.5%; temporal and structural features equally useful.

Socially Shared Information



Comment Threads

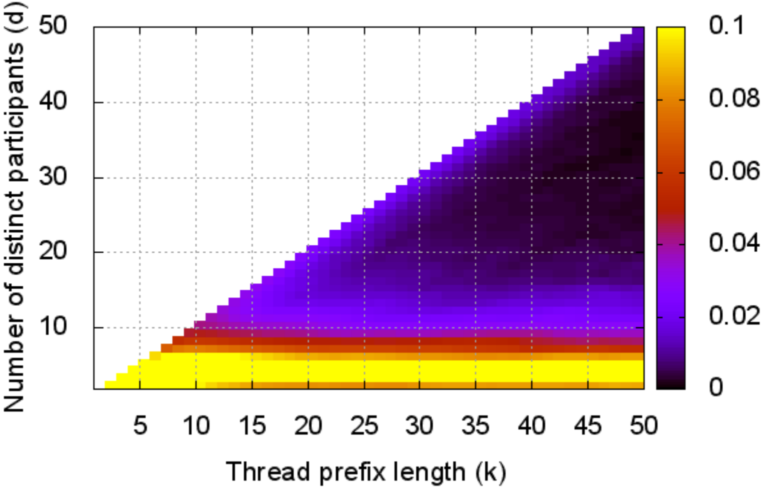
Different mode: users post; friends comment.

- Can we predict the eventual length of a comment thread?

Multiple ways for a post to be long

- $\Delta_k(d) =$ fraction of length-k threads with d distinct commenters.
- A new problem: re-entry prediction.

Number of distinct commenters

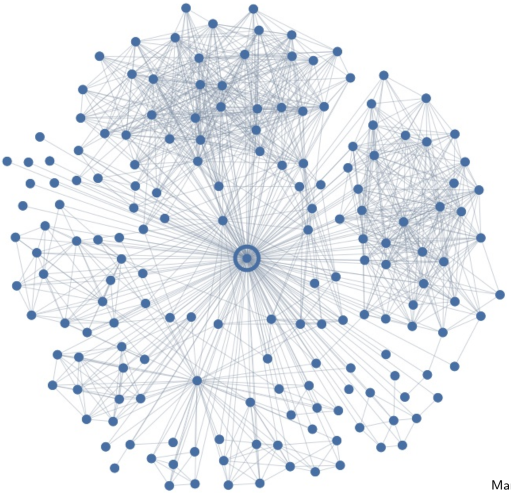


Comment Threads

Not yet modeled: Identity of poster has clear importance.

- Typical FB user writes 60-70% of comments to $\approx$ 15 people. [Backstrom-Bakshy-Kleinberg-Lento-Rosenn 2011]

Network Neighborhoods



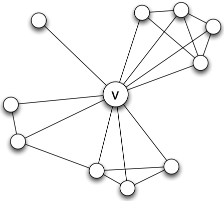
Marlow-Byron-Lento-Rosenn 2009

One person's network neighborhood:

- The "input" for their experience in a social-networking system.

Finding Significant People

Given a person's network neighborhood, can we identify their most significant social ties?

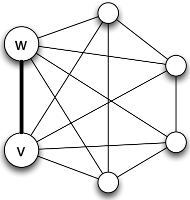


Theories of strong and weak ties [Granovetter 1973, 1985].

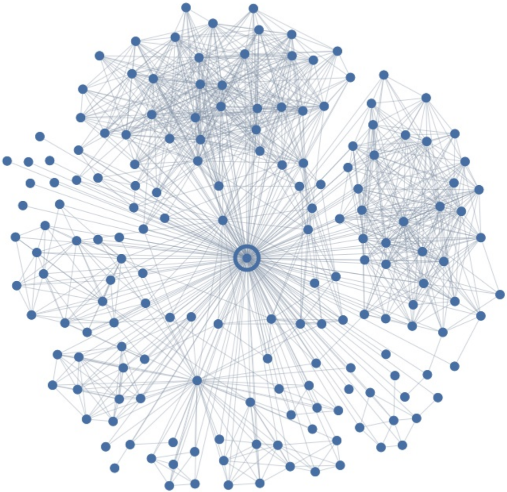
- Embeddedness: # of mutual friends shared by e's endpoints.

If an edge is highly embedded, it is likely to be a stronger tie.

- Rank neighbors by embeddedness?

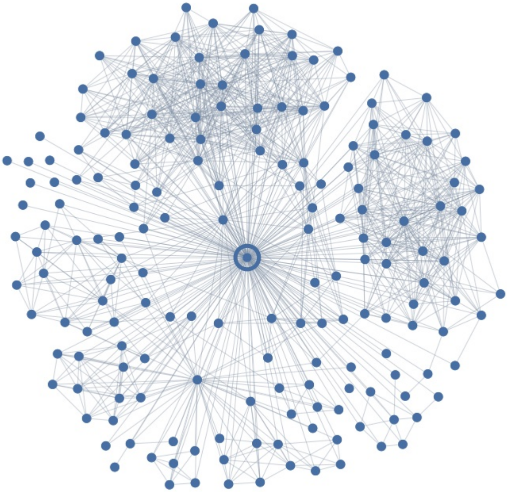


Network structure via neighborhoods



In practice: embeddedness finds many nodes from the largest cluster.

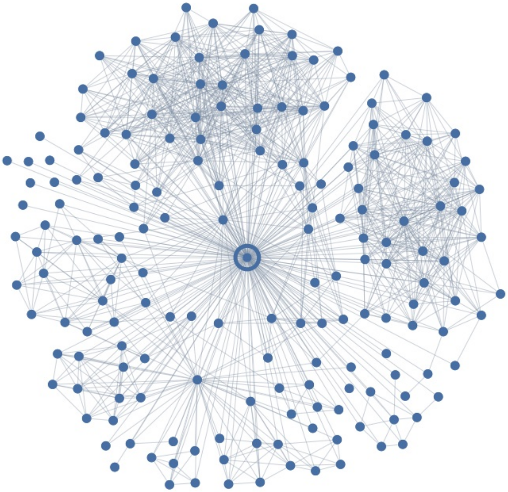
Network structure via neighborhoods



In practice: embeddedness finds many nodes from the largest cluster.

- Often this is a large collection of co-workers or college alumni friends. Compare: node in lower left — the spouse.

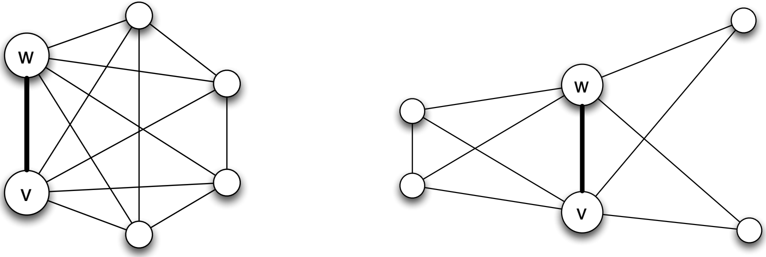
Network structure via neighborhoods



In practice: embeddedness finds many nodes from the largest cluster.

- Motivating question: Given a Facebook user in a relationship, find their partner just from network structure [Backstrom-Kleinberg 2014]

Alternatives to Embeddedness

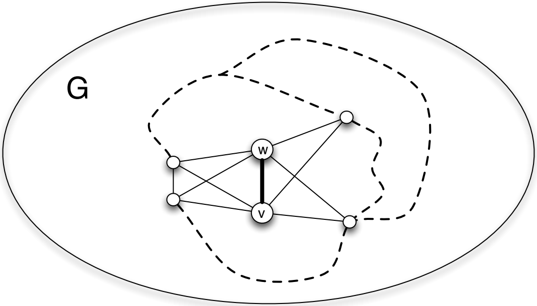


Instead of just counting mutual friends, look at their structure.

- How well connected are the common endpoints of edge e?
- If not well connected, suggests something about v-w relationship.
- v-w cannot be easily "explained" by any one social focus.

Type of bridging/brokerage role [Granovetter 73, Burt 92, Watts 99] but played jointly by v and w, and implying a form of tie strength.

Dispersion



C_{vw} = common neighbors of v and w.

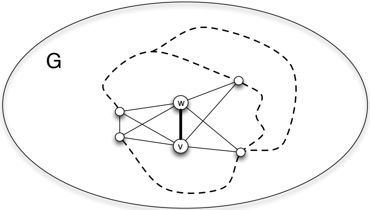
Sum of distances between pairs in C_{vw}, after deleting v and w:

$$\sum_{s,t \in C_{vw}} d_{G - \{v,w\}}(s, t).$$

The dispersion of edge (v, w) with respect to distance function d.

- Based on a 0-1-valued metric, normalized by $|C_{vw}|$.

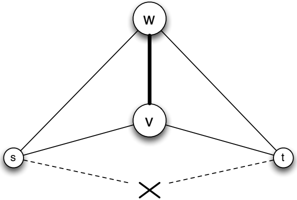
Can use many possible distance functions d when summing over pairs of mutual neighbors.



- $d(s,t) = \begin{cases} 0 \text{ if } (s,t) \text{ is an edge} \\ 1 \text{ otherwise} \end{cases}$
- $d(s,t) = \begin{cases} 0 \text{ if shortest } s\text{-}t \text{ path avoiding } v, w \text{ has } \leq k \text{ edges} \\ 1 \text{ otherwise} \end{cases}$
- Many other choices for d based on community detection, brokerage measures, spring embedding, ...

Can also normalize the dispersion:

$$\frac{\mathit{dispersion}(v,w)}{(\#\ \text{mutual nbrs})^{\alpha}}.$$



- Searching over choices of k, α shows $k = 2$ and $\alpha = 1$ nearly optimal.
- A slight improvement if we apply this recursively (details omitted here ...)

Evaluating the Methods

For evaluation, use 1.3 million Facebook users who:

- Declare a relationship partner in their profile (symmetric).
- Have between 50 and 2000 friends.
- Are at least 20 years old.

For each user v, rank all friends w by competing metrics:

- Embeddedness of v-w edge.
- Dispersion of v-w edge.
- Number of photos in which v and w are both tagged.
- Number of times v viewed w's profile in last 90 days.

For what fraction of all users v is the top-ranked w the relationship partner?

type	embed	dispersion	photo	profile view
all	0.247	0.506		

Notes:

Embeddedness vs. dispersion

type	embed	dispersion	photo	profile view
all	0.247	0.506	0.415	0.301

Notes:

Embeddedness vs. dispersion

Structural vs. activity-based

type	embed	dispersion	photo	profile view
all	0.247	0.506	0.415	0.301
married	0.321	0.607	0.449	0.210
relationship	0.132	0.344	0.347	0.441

Notes:

Embeddedness vs. dispersion

Structural vs. activity-based

Married vs. in a relationship

type	embed	dispersion	photo	profile view
all	0.247	0.506	0.415	0.301
married	0.321	0.607	0.449	0.210
married (female)	0.296	0.551	0.391	0.202
married (male)	0.347	0.667	0.511	0.220
relationship	0.132	0.344	0.347	0.441
relationship (female)	0.139	0.316	0.290	0.467
relationship (male)	0.125	0.369	0.399	0.418

Notes:

- Embeddedness vs. dispersion
- Structural vs. activity-based
- Married vs. in a relationship
- Female vs. male

type	embed	dispersion	photo	profile view
all	0.247	0.506	0.415	0.301
married	0.321	0.607	0.449	0.210
married (female)	0.296	0.551	0.391	0.202
married (male)	0.347	0.667	0.511	0.220
relationship	0.132	0.344	0.347	0.441
relationship (female)	0.139	0.316	0.290	0.467
relationship (male)	0.125	0.369	0.399	0.418

Notes:

Embeddedness vs. dispersion

Structural vs. activity-based

Married vs. in a relationship

Female vs. male

Combining all via machine learning: 0.716 married, 0.682 relationship

type	embed	dispersion	photo	profile view
all	0.247	0.506	0.415	0.301
married	0.321	0.607	0.449	0.210
married (female)	0.296	0.551	0.391	0.202
married (male)	0.347	0.667	0.511	0.220
relationship	0.132	0.344	0.347	0.441
relationship (female)	0.139	0.316	0.290	0.467
relationship (male)	0.125	0.369	0.399	0.418

Notes:

- Embeddedness vs. dispersion
- Structural vs. activity-based
- Married vs. in a relationship
- Female vs. male
- Combining all via machine learning: 0.716 married, 0.682 relationship
- Approx 34-38% of dispersion's incorrect guesses are family members.

type	embed	dispersion	photo	profile view
all	0.247	0.506	0.415	0.301
married	0.321	0.607	0.449	0.210
married (female)	0.296	0.551	0.391	0.202
married (male)	0.347	0.667	0.511	0.220
relationship	0.132	0.344	0.347	0.441
relationship (female)	0.139	0.316	0.290	0.467
relationship (male)	0.125	0.369	0.399	0.418

Notes:

Embeddedness vs. dispersion

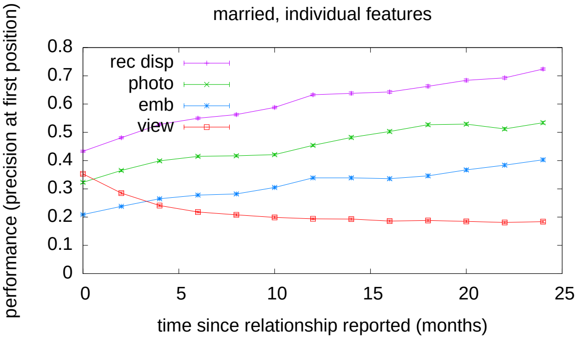
Structural vs. activity-based

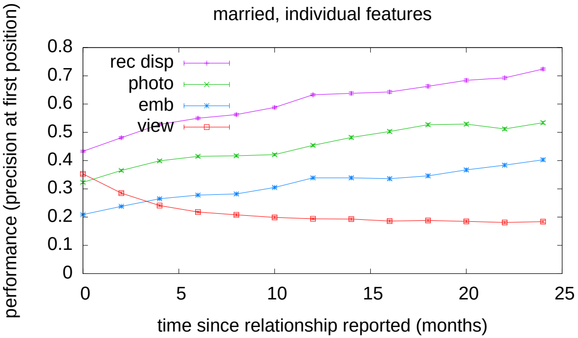
Married vs. in a relationship

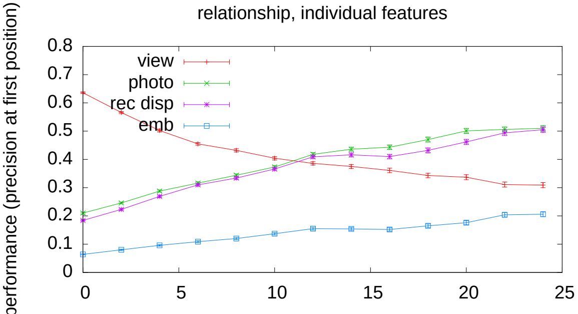
Female vs. male

Combining all via machine learning: 0.716 married, 0.682 relationship

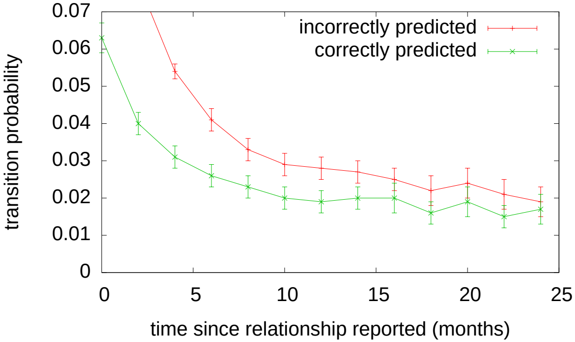
Approx 34-38% of dispersion's incorrect guesses are family members.


married, individual features
performance (precision at first position)
time since relationship reported (months)
rec disp
photo
emb
view
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
5
10
15
20
25


married, individual features
performance (precision at first position)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
rec disp
photo
emb
view
0
5
10
15
20
25
time since relationship reported (months)


relationship, individual features
performance (precision at first position)
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
view
photo
rec disp
emb
0
5
10
15
20
25

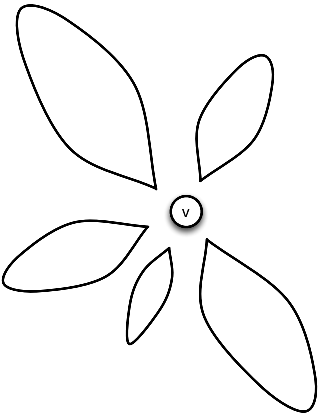
Probability a relationship ends



Probability a user transitions to 'single' status in next 60 days.

- Relationships where dispersion is correct vs. incorrect.
- Separately over relationships in 2-month age ranges.

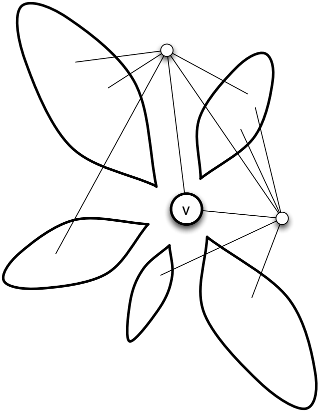
A General Structure for Network Neighborhoods



A schematic picture for a node's neighborhood:

A constant number of homogeneous clusters.

A General Structure for Network Neighborhoods



A schematic picture for a node's neighborhood:

A constant number of homogeneous clusters.

Plus a constant number of nodes that defy classification.

Designed Social Systems

Computational challenges in managing on-line social systems.

- Algorithmically identifying and filtering content as it flows in the network.
- Network neighborhoods as central structures [Ugander et al 2012, 2013]
- Incentives to propagate information: e.g.
  Query incentive networks [Kleinberg-Raghavan 2005],
  DARPA Network Challenge [Pickard et al 2011],
  Bitcoin [Babaioff et al 2012].
- Integration with language analysis [Danescu-Niculescu-Mizil et al 2011].