Link formation mechanisms

Introduction to Network Science

Carlos Castillo

Topic 07

Universitat Pompeu Fabra Barcelona

Sources

- Albert László Barabási: Network Science. Cambridge University Press, 2016. Ch 07
- Networks, Crowds, and Markets Ch 03 and 04
- C. Castillo: Link prediction slides 2016

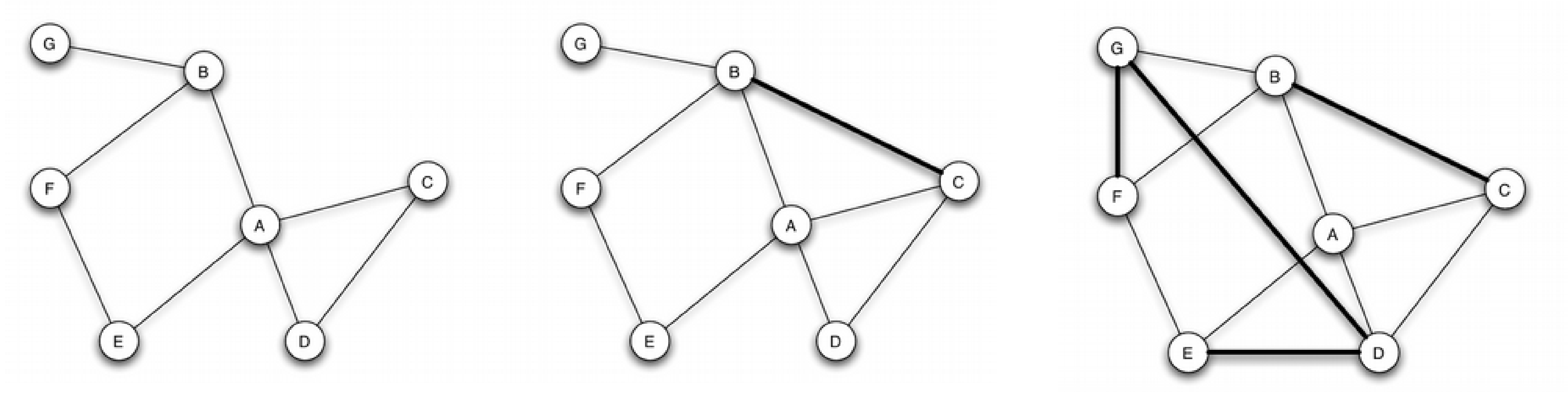
Link formation is contextual

- It is affected by existing links
  - e.g., Triadic closure
  - It is also affected by content sharing
- It is affected by node affinity/similarity
  - e.g., Similar characteristics
  - e.g., Similar degree

Triadic closure

- If two nodes in a network …
  - are not connected,
  - but have connections in common,
- … there is a larger probability that they will form a connection in the future

Triadic closure example



Which edges are triadic closures?

Possible mechanisms for triadic closure

- Opportunity: A meets B and C often, eventually B and C will meet
- Trust: A trusts B, A trusts C, B can trust C
- Incentive: if A is friend with B and C, but B and C are not friends, there is tension/stress
  - Teenage girls with low clustering coefficient more likely to contemplate suicide

Peter Bearman and James Moody. Suicide and friendships among American adolescents. American Journal of Public Health, 94(1):89–95, 2004.

Possible mechanisms for triadic closure

- Opportunity: A meets B and C often, eventually B and C will meet
- Trust: A trusts B, A trusts C, B can trust C
- Incentive: if A is friend with B and C, but B and C are not friends, there is tension/stress
- Triadic closures can happen **even if B and C don’t know that they have a friend in common!**

Strong/Weak Triadic Closure

- Triadic closure can be observed in weighted graphs
- If A-B and A-C have a strong connection, then strong triadic closure is violated if B-C have a weak connection or no connection at all

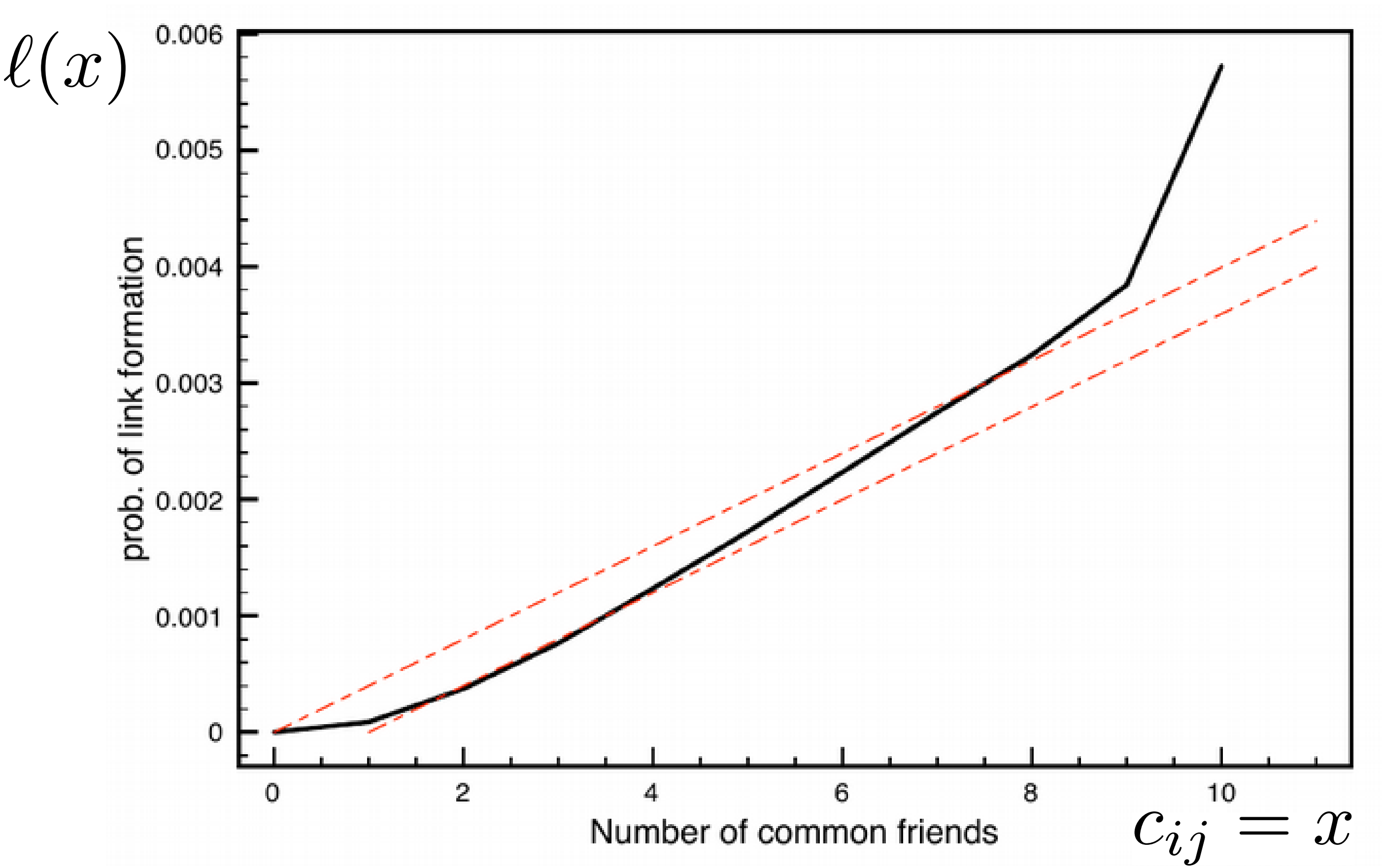
Triadic closure and common neighbors

- Let c_{ij} be the number of neighbors in common between nodes i and j
- Suppose we take two snapshots: E_{t_0}, E_{t_1}
- We want to study the function

$$\ell(x) = Pr[(i,j) \in E_{t_1} | (i,j) \notin E_{t_0} \wedge c_{ij} = x]$$

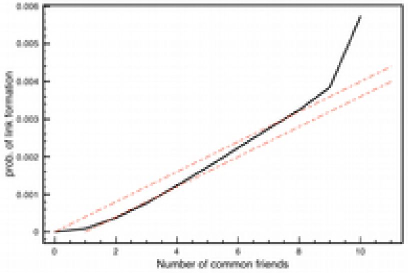
How should $\ell(x)$ be with respect to x ?

Study in an e-mail dataset



Kossinets, G., & Watts, D. J. (2006). Empirical analysis of an evolving social network. science, 311(5757), 88-90.

Details [Kossinets & Watts]



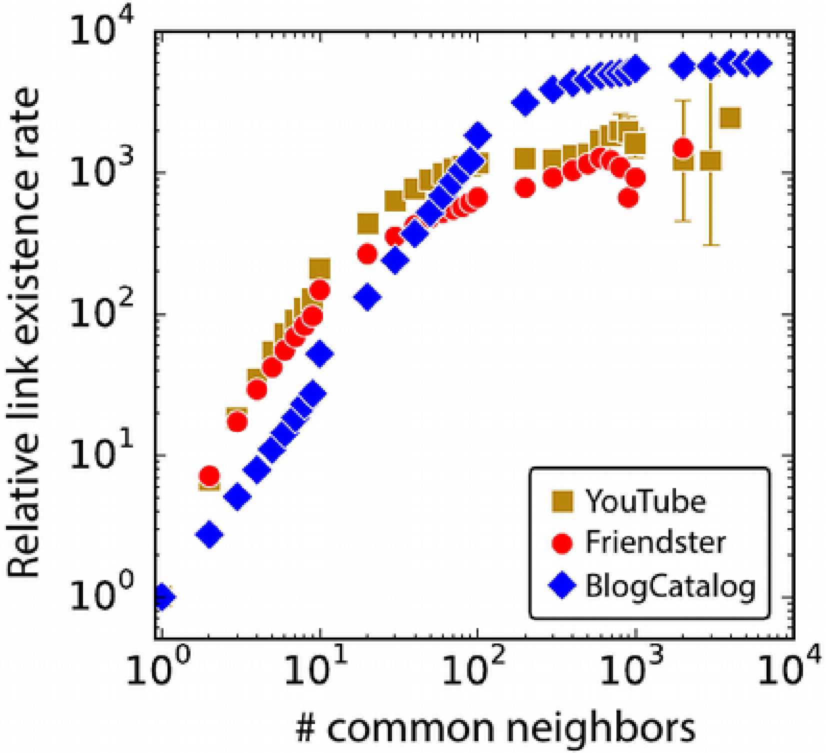
- Dataset:
  - Anonymized e-mails between 22,000 students
  - Edge(i,j) is present if the users i and j have exchanged at least 1 e-mail in the past 60 days
  - One "snapshot" per day
- Curve shown is an average
  - Multiple pairs of snapshots separated by one day

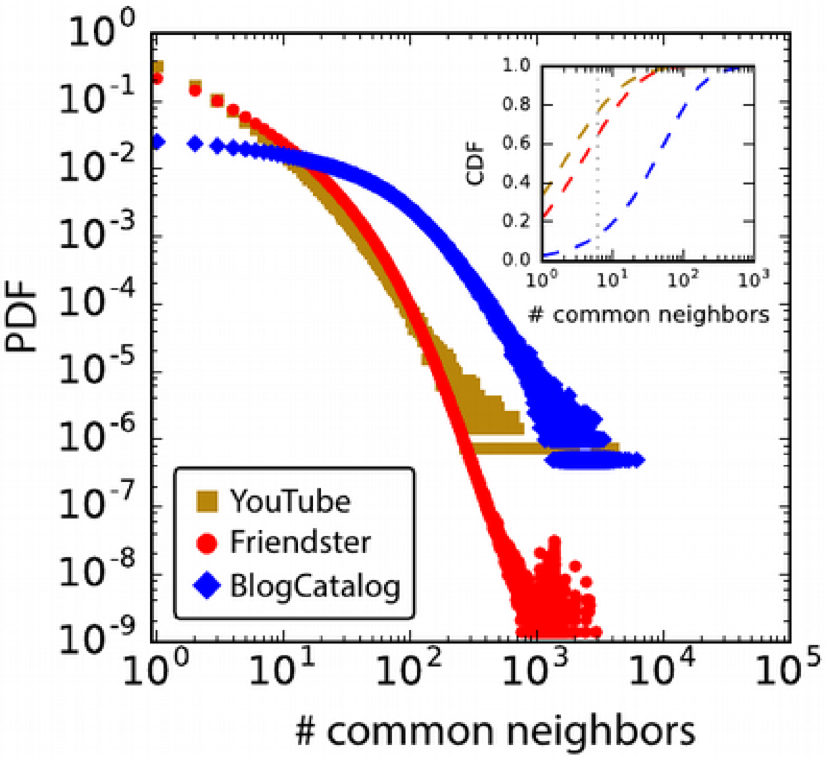
Simple model for $\ell(x)$

- Node i is not connected to node j
- Between t_0 and t_1 node i sees all the common friends s/he has with node j
- Each time there is a small chance *p* they will introduce node i to j

Write $\ell(x)$ as a function of x and p

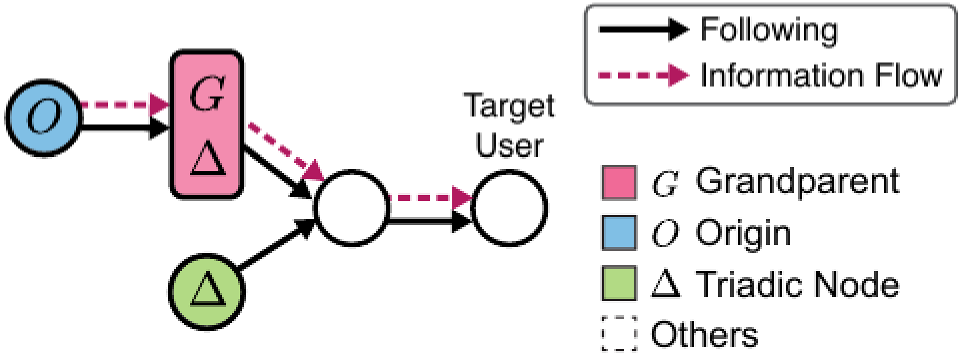
Evidence from other networks

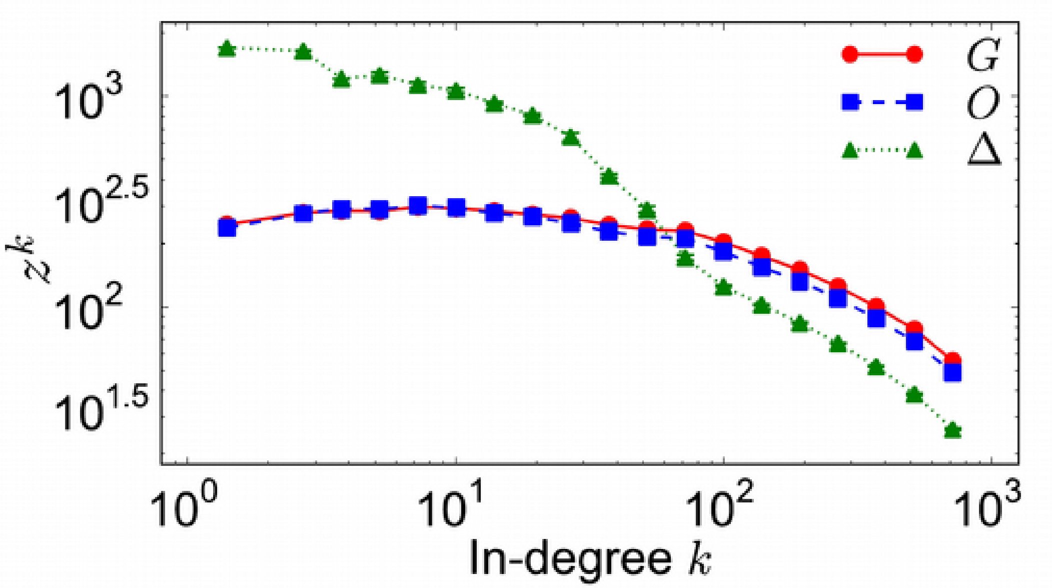




Dong, Y., Johnson, R. A., Xu, J., & Chawla, N. V. (2017, August). Structural diversity and homophily: A study across more than one hundred big networks. In Proc. KDD. ACM.

Link probability and content sharing



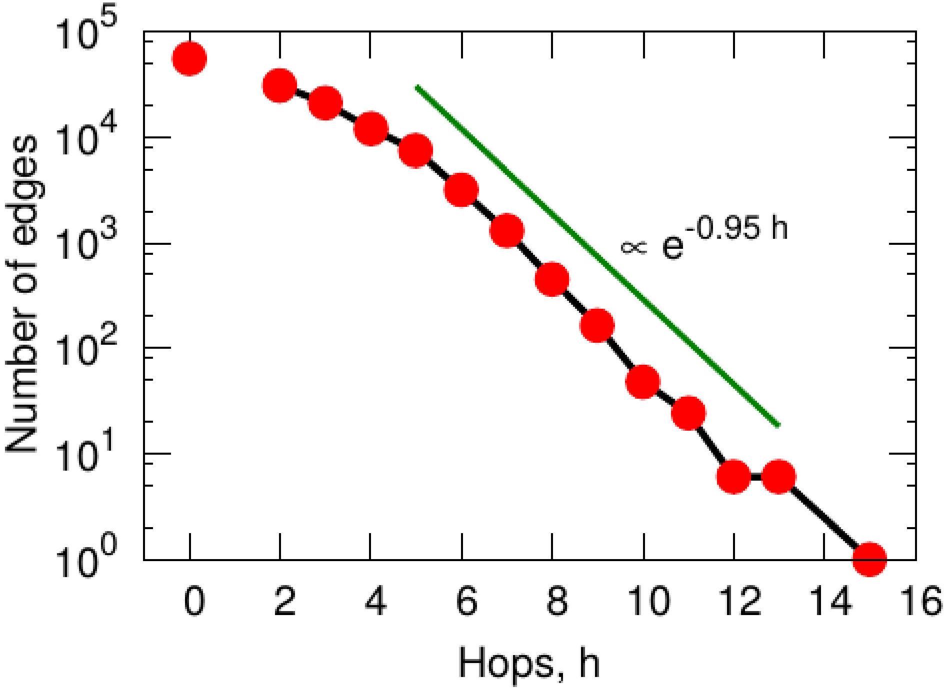


Triadic closures are affected by content consumption, and are more likely to happen with nodes from which I have received content (posts or re-posts)

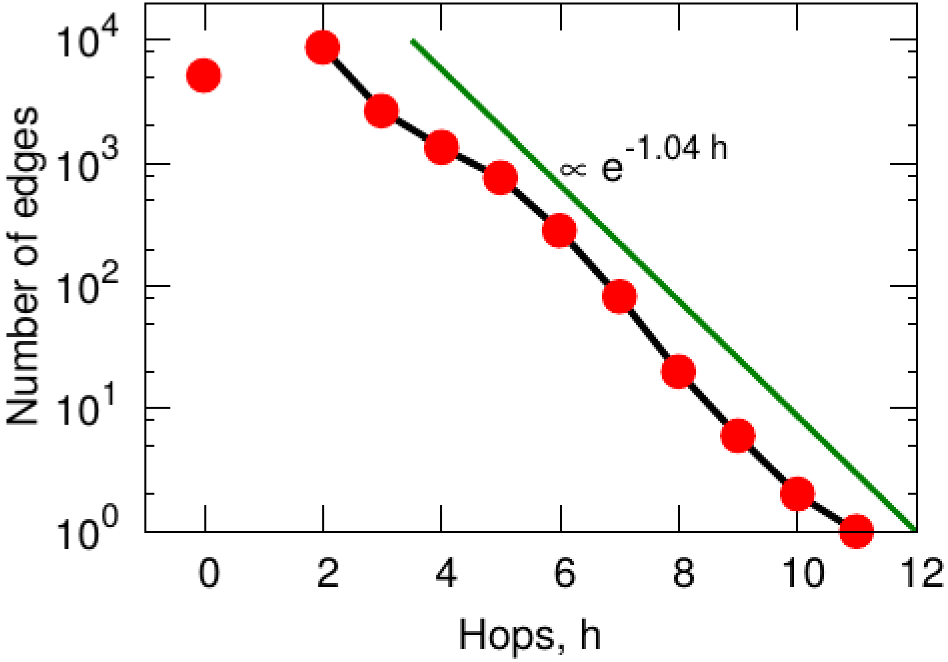
Weng, L., Ratkiewicz, J., Perra, N., Gonçalves, B., Castillo, C., Bonchi, F., Flammini, A. (2013). The role of information diffusion in the evolution of social networks. In KDD. ACM.

Linking probability and distance ("hitting time")

Yahoo! Answers



LinkedIn



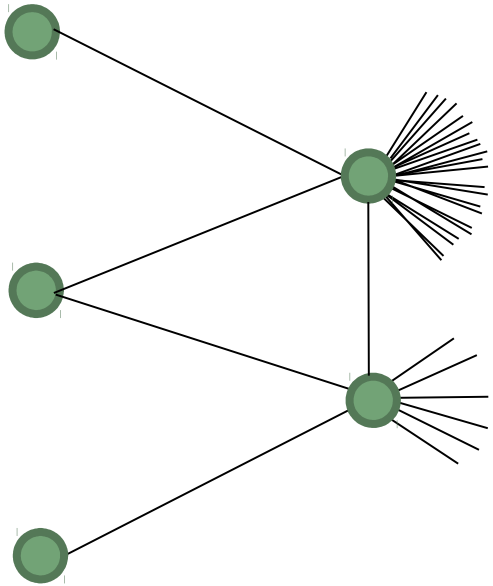
Leskovec, J., Backstrom, L., Kumar, R., & Tomkins, A. (2008). Microscopic evolution of social networks. In Proc. KDD

Useful scores to predict linking

- Jaccard similarity $score(i,j) = \frac{c_{ij}}{|\Gamma(i) \cup \Gamma(j)|}$
- Adar-Adamic score $score(i,j) = \sum_{z \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log(k_z)}$

The idea is to avoid this

Understanding the Adamic-Adar score



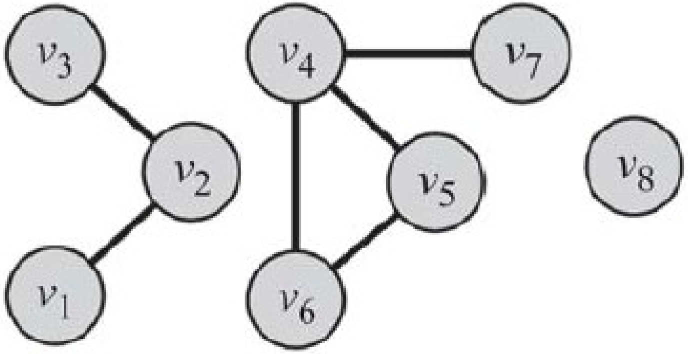
1000 followers

Prolific common friends
Weaker evidence

8 followers

Less prolific
Stronger evidence

Would you recommend (1,3) or (5,7)?

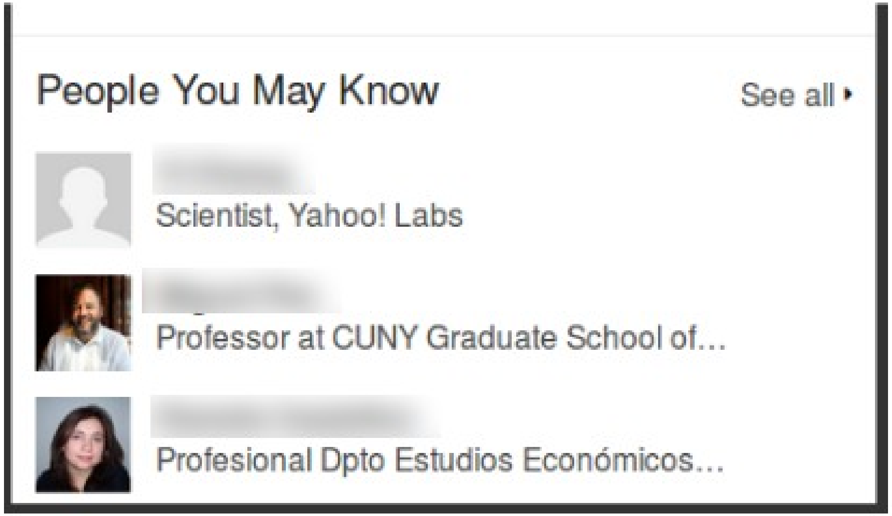


Compare using:

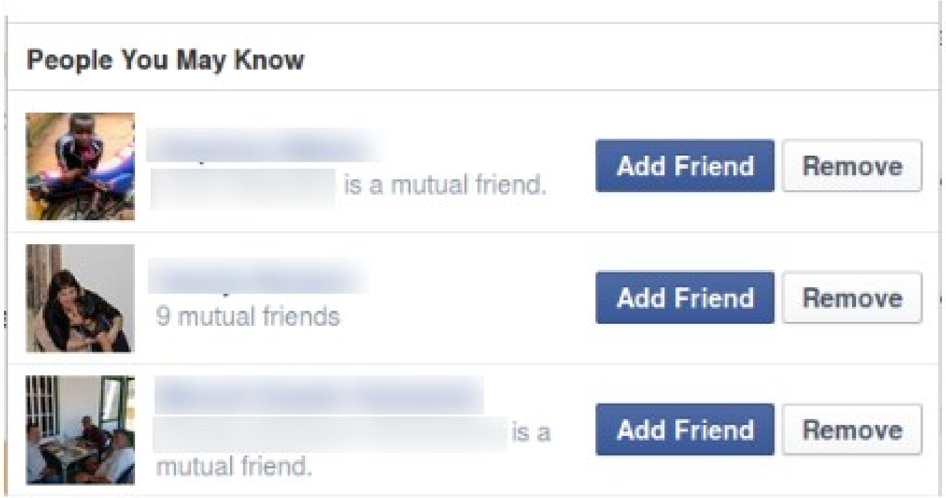
- *Number of common neighbors*
- *Jaccard coefficient*
- *Adamic-Adar*
  (+1 to denominator if needed)

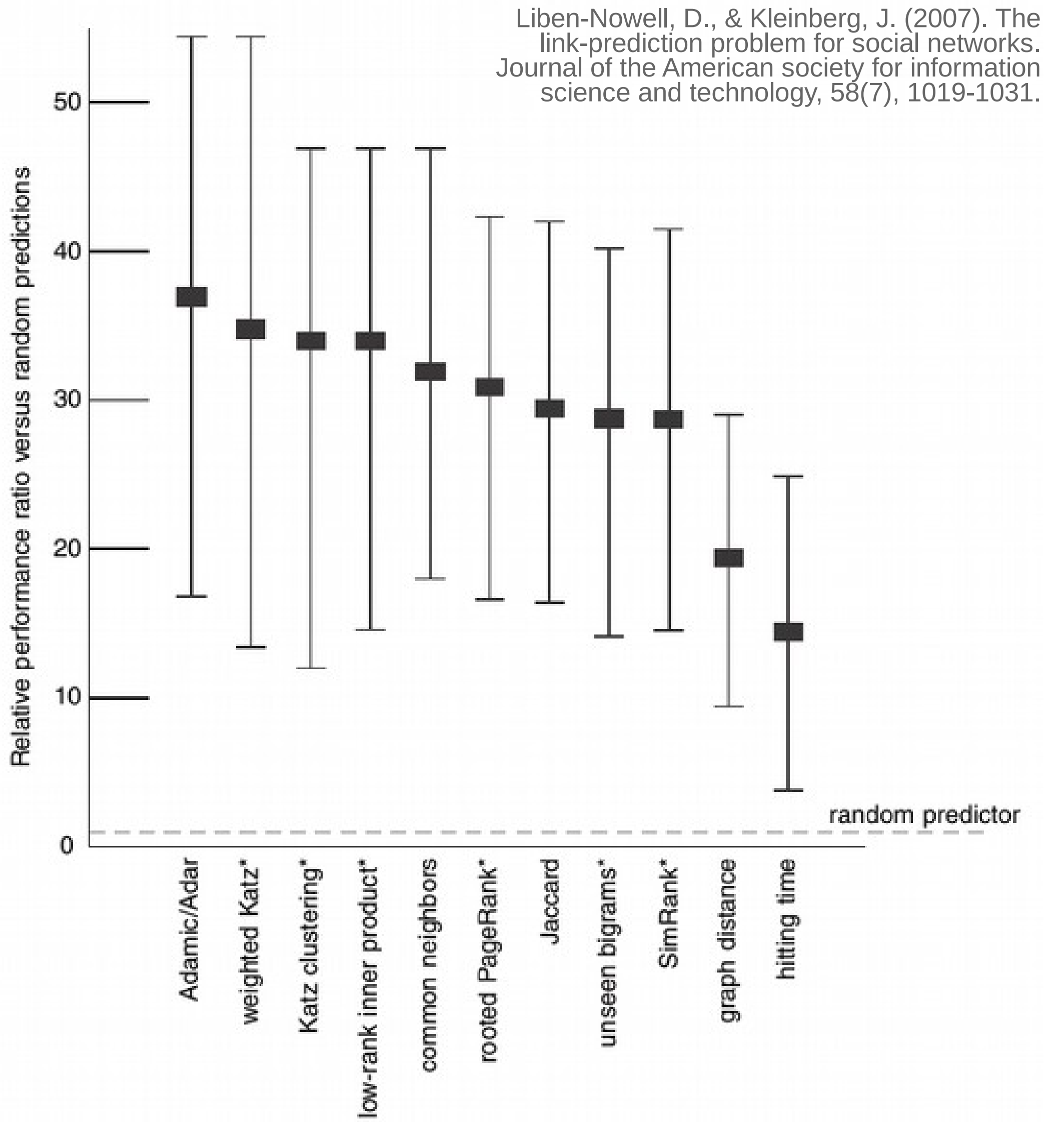
Application: link prediction

LinkedIn



facebook





Comparison

This is a hugely imbalanced problem, imagine all the friends you could but did NOT make last year!

Accuracy is very low unless you play it safe … and then it's not very useful

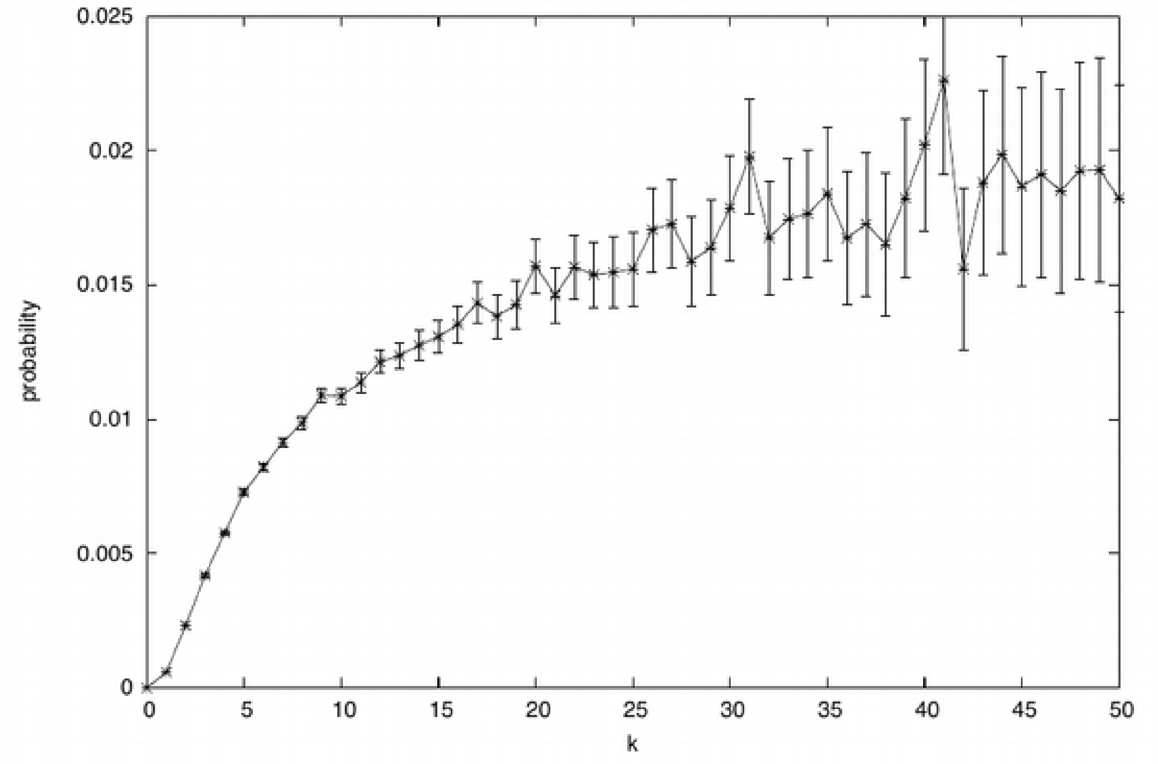
Community membership

Community membership prediction

- Why do users join communities?
- We can observe users who join communities and determine the factors that are common among them
- We require a population of users, a community C, and community membership information (i.e., users who are members of C).
  - To distinguish between users who have already joined the community and those who are now joining it, we need community memberships at two different times t_1, t_2

Peer influence

Probability of joining an online community given k friends are already members

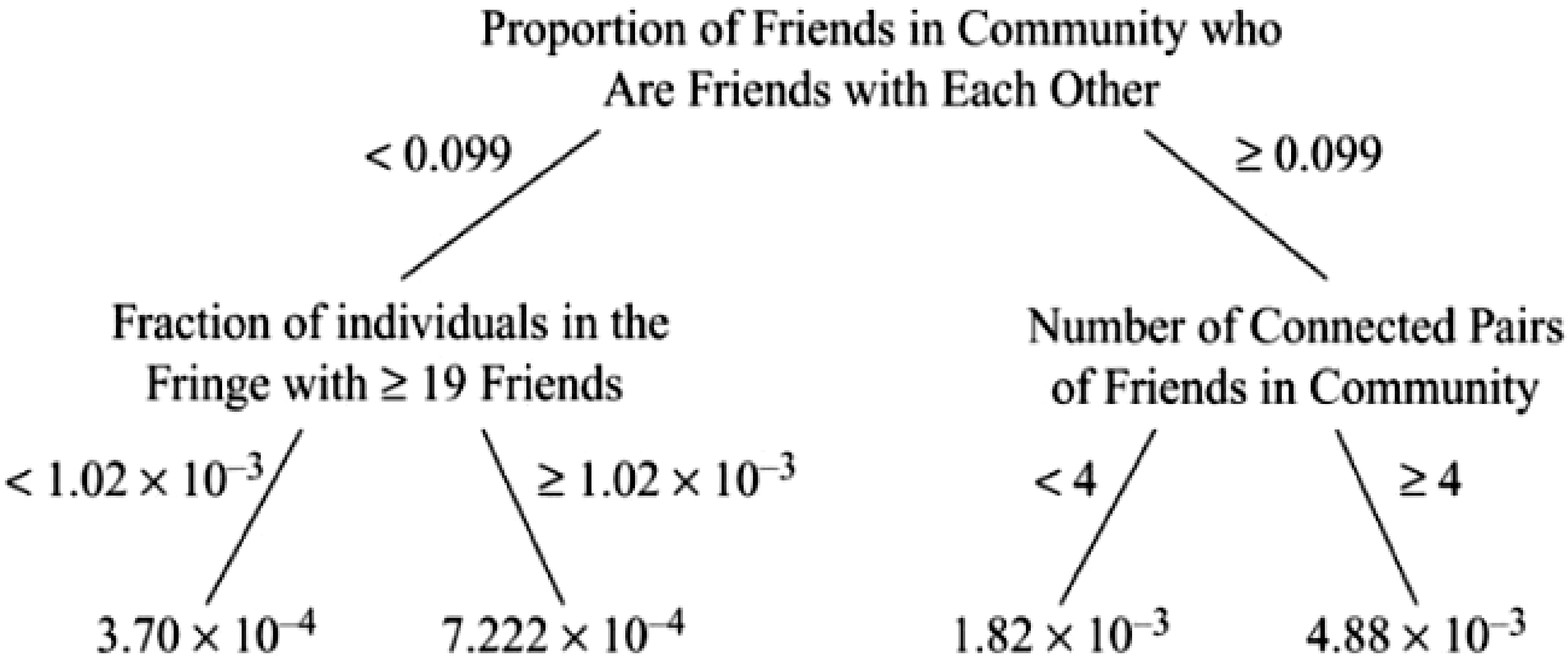


Lars Backstrom, Dan Huttenlocher, Jon Kleinberg, and Xiangyang Lan. Group formation in large social networks: Membership, growth, and evolution. In Proc. KDD.

Idea: supervised learning

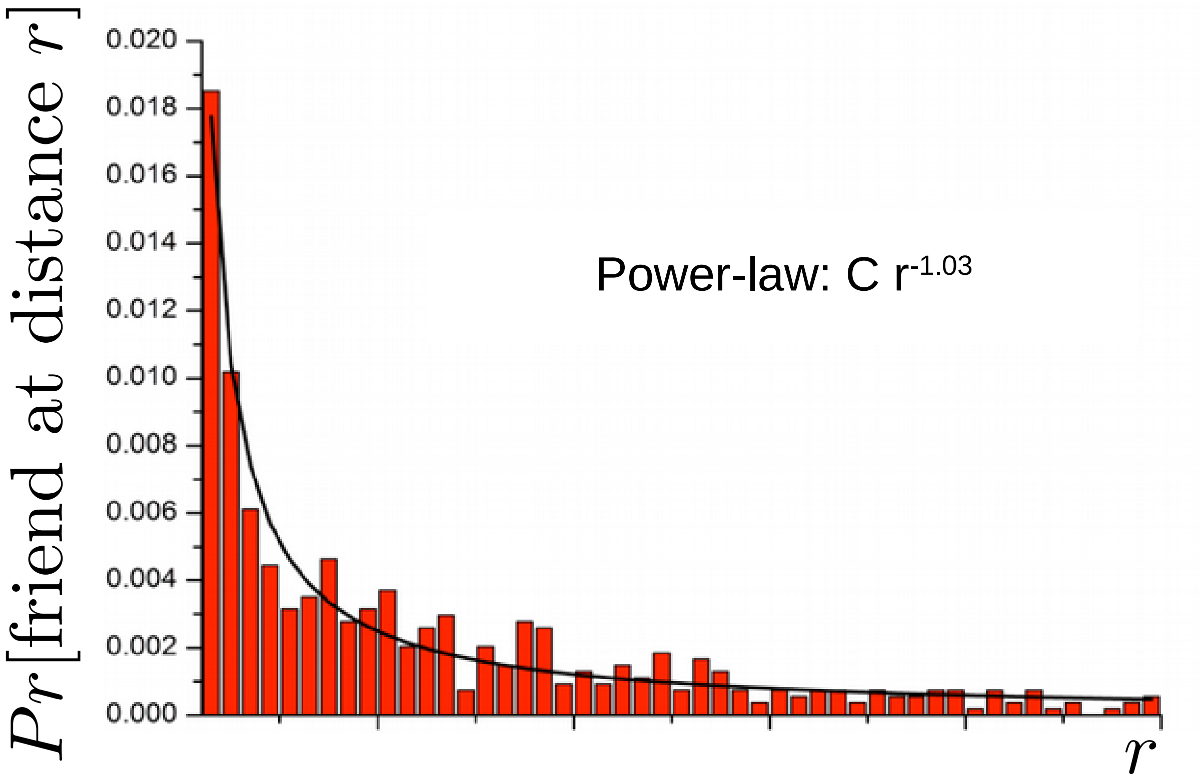
Feature Set	Feature
Features related to the community, C. (Edges between only members of the community are $E_C \subseteq E$.)	Number of members ($\|C\|$). Number of individuals with a friend in C (the *fringe* of C) . Number of edges with one end in the community and the other in the fringe. Number of edges with both ends in the community, $\|E_C\|$. The number of open triads: $\|\{(u,v,w)\|(u,v) \in E_C \wedge (v,w) \in E_C \wedge (u,w) \notin E_C \wedge u \neq w\}\|$. The number of closed triads: $\|\{(u,v,w)\|(u,v) \in E_C \wedge (v,w) \in E_C \wedge (u,w) \in E_C\}\|$. The ratio of closed to open triads. The fraction of individuals in the fringe with at least k friends in the community for $2 \leq k \leq 19$. The number of posts and responses made by members of the community. The number of members of the community with at least one post or response. The number of responses per post.
Features related to an individual u and her set S of friends in community C.	Number of friends in community ($\|S\|$). Number of adjacent pairs in S ($\|\{(u,v)\|u,v \in S \wedge (u,v) \in E_C\}\|$). Number of pairs in S connected via a path in E_C. Average distance between friends connected via a path in E_C. Number of community members reachable from S using edges in E_C. Average distance from S to reachable community members using edges in E_C. The number of posts and response made by individuals in S. The number of individuals in S with at least 1 post or response.

Example regression tree



Networks and geography

Distance is not dead (The world is not "flat")

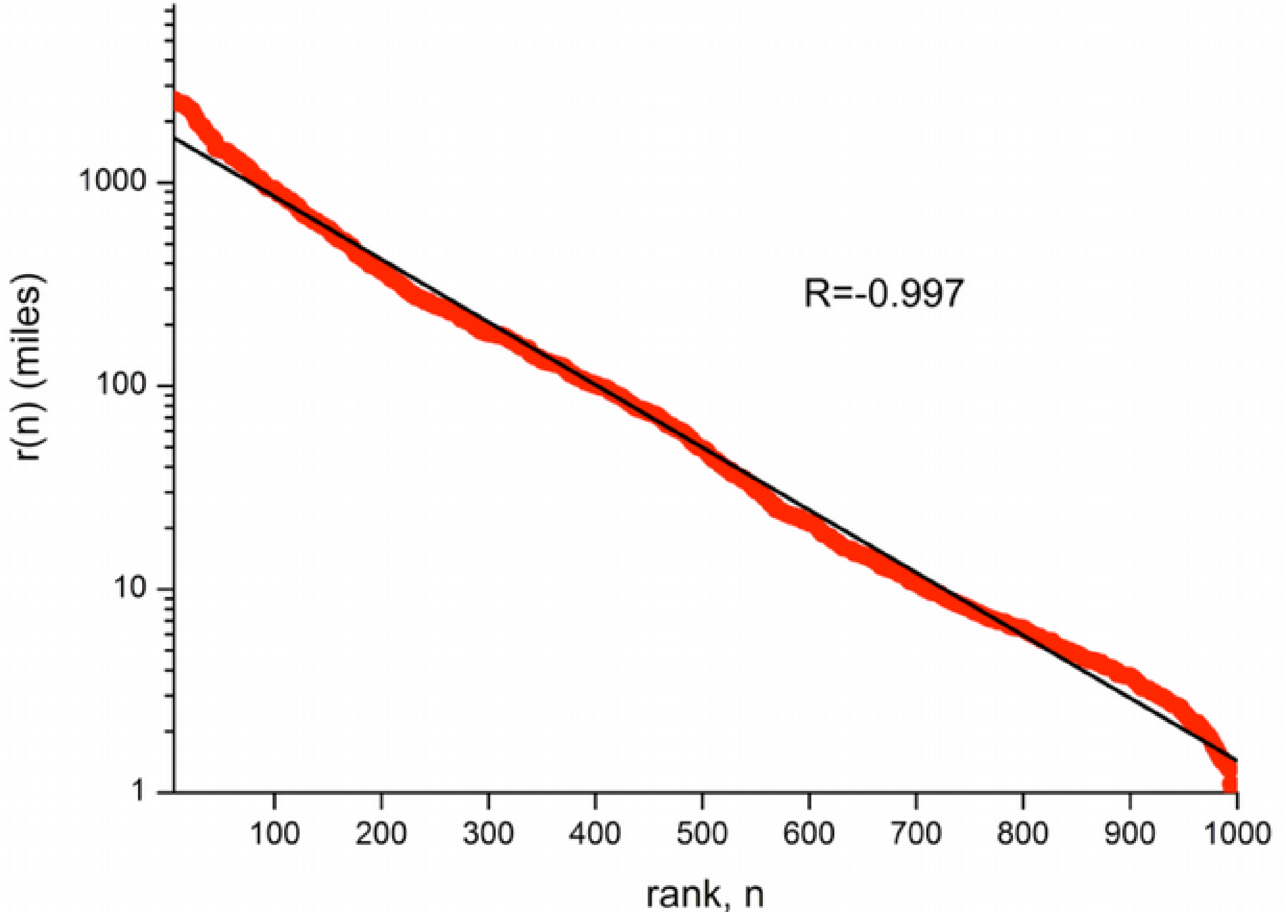


The probability of being friends decreases rapidly with distance, but … you can still have friends far away

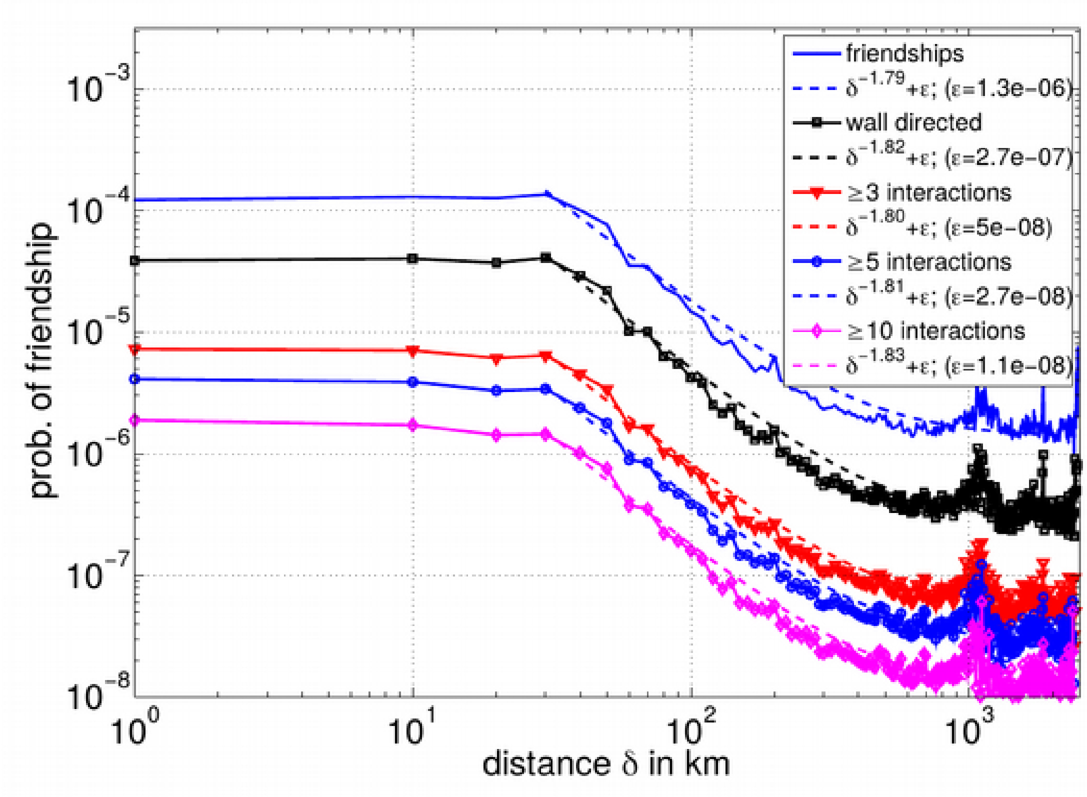
(follows power-law, not exponential decay)

Goldenberg, J., & Levy, M. (2009). Distance is not dead: Social interaction and geographical distance in the internet era. arXiv preprint arXiv:0906.3202.

Sorting friends from most distant (geographically) to closest one



Tie strength and geographical distance (data from Tuenti Spain)

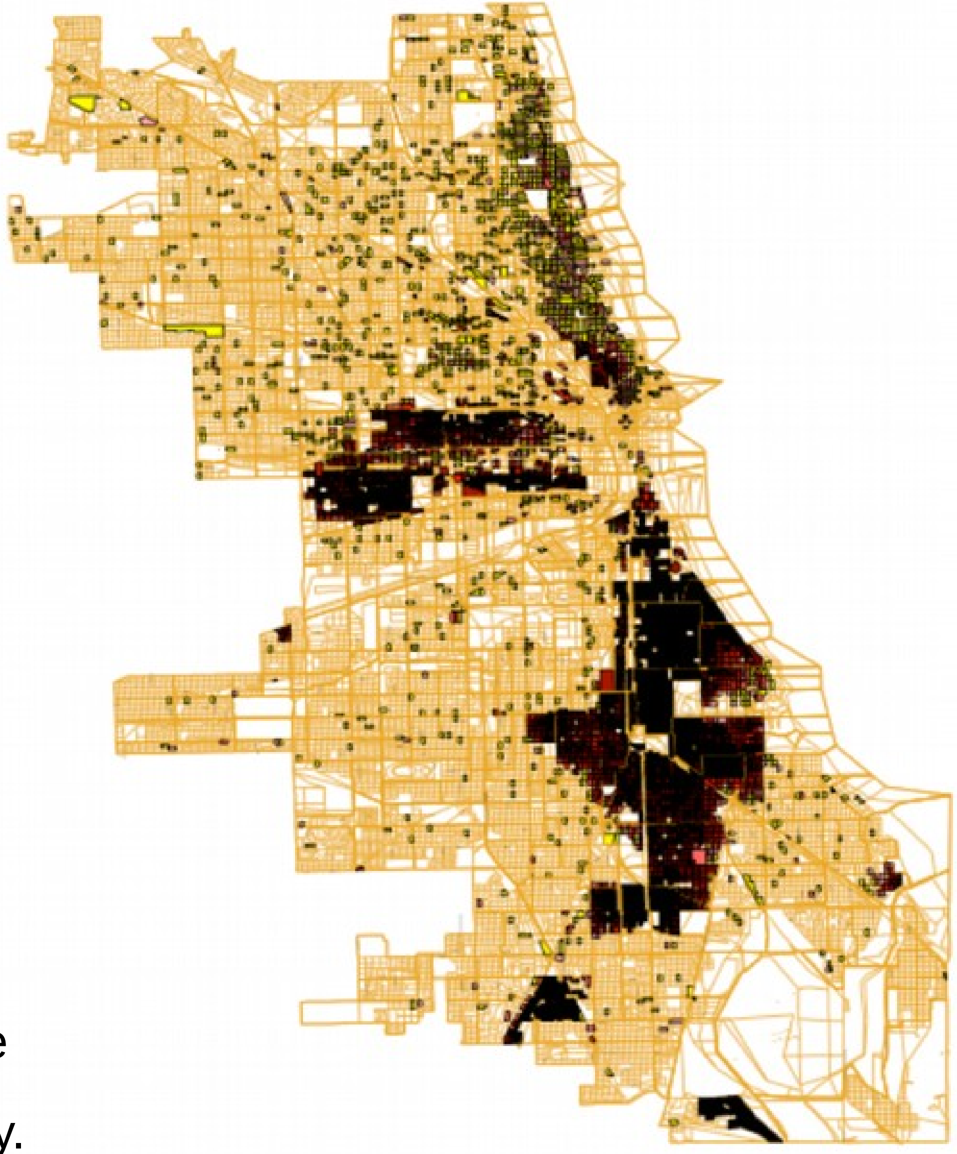


In Tuenti's data there is a clear drop at 30km of distance

Laniado, David, Yana Volkovich, Salvatore Scellato, Cecilia Mascolo, and Andreas Kaltenbrunner. "The impact of geographic distance on online social interactions." Information Systems Frontiers (2017): 1-16.

Geographical segregation

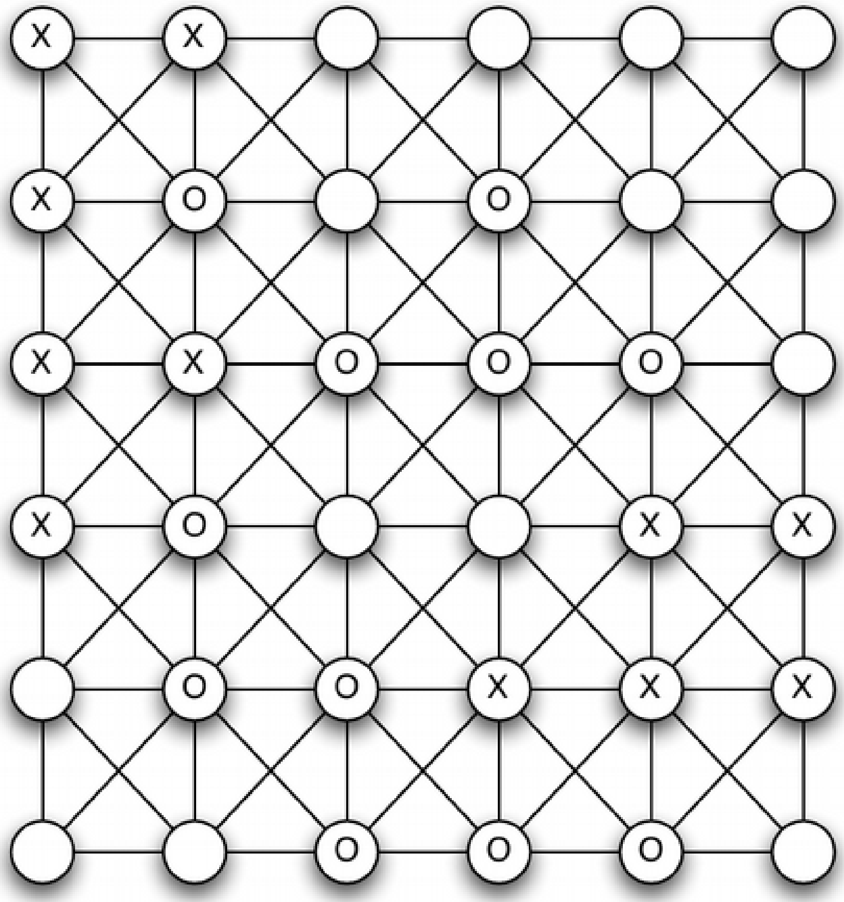
In this map of Chicago (US) in 1960, brown/black areas have majority African-American populations



Möbius, M. M., & Rosenblat, T. S. (2001). The process of ghetto formation: evidence from Chicago. Technical Report, Harvard University.

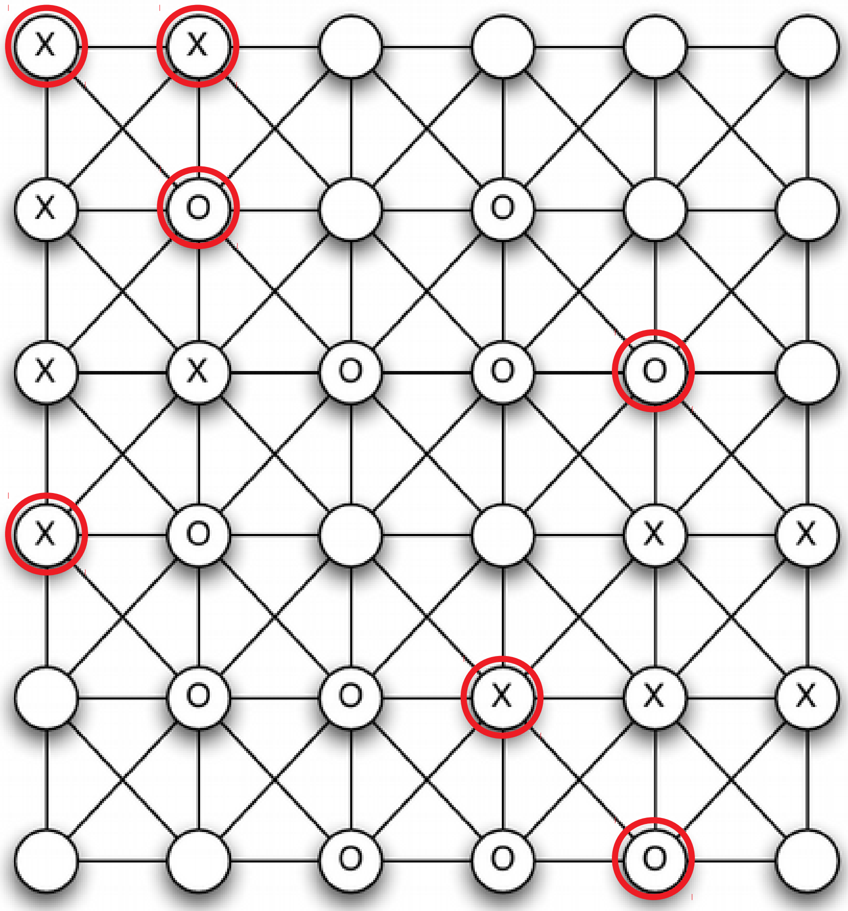
Schelling Model

- Two types of people: O, X
- Living in a lattice
  (8 neighbors, except borders)
- You are **satisfied** if you have at least t neighbors of your own kind
- Otherwise you are **unsatisfied** and you must move to an adjacent cell



Unsatisfied nodes (t=3)

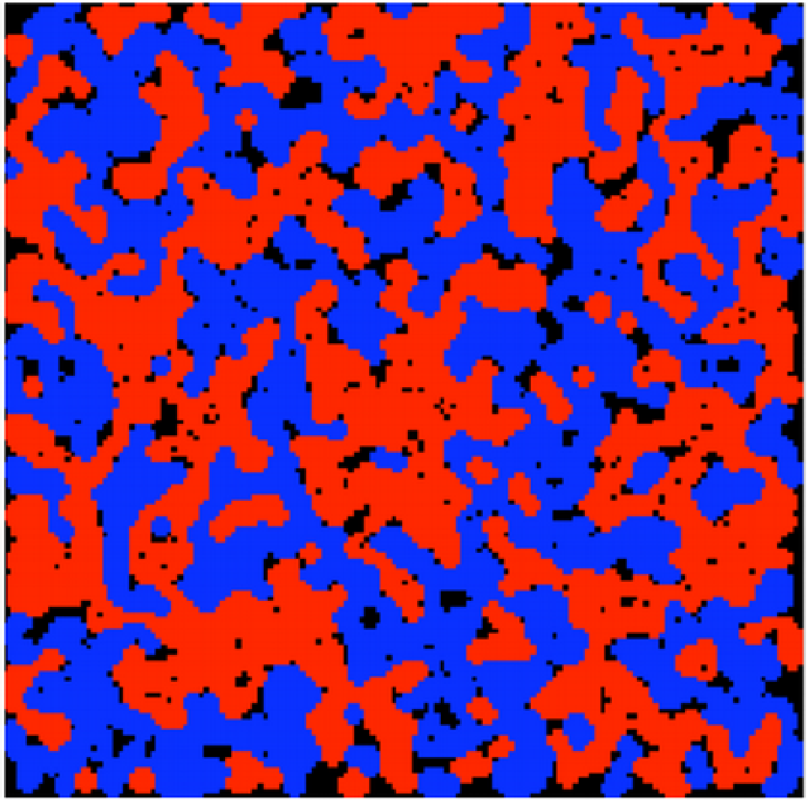
- Two types of people: O, X
- Living in a lattice
  (8 neighbors, except borders)
- You are **satisfied** if you have at least t neighbors of your own kind
- Otherwise you are **unsatisfied** and you must move to an adjacent cell

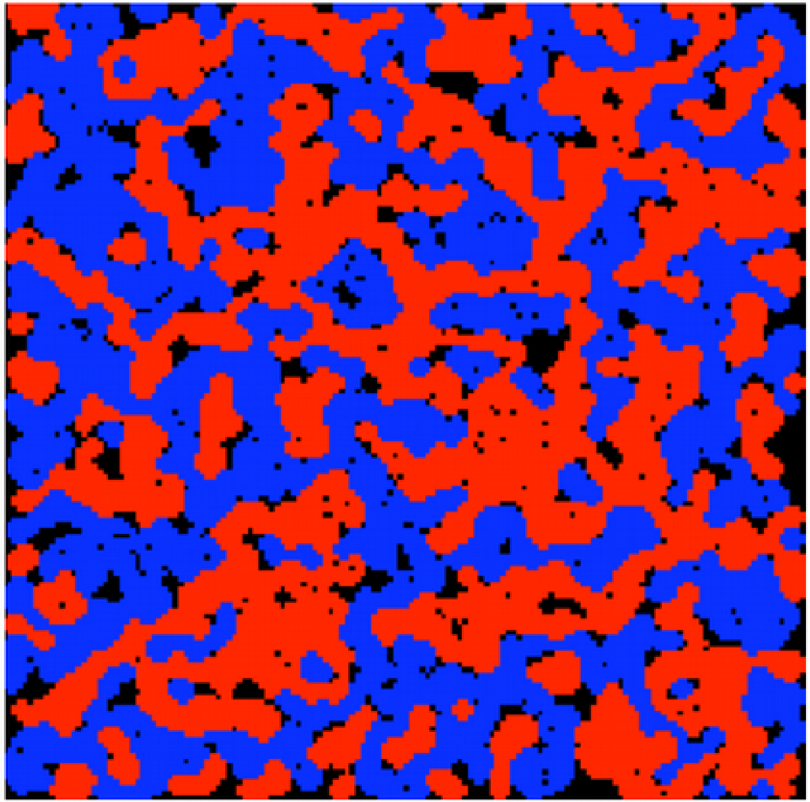


Details

- The process proceeds in rounds
- Sometimes nodes cannot be satisfied
  - They can be randomly placed or left in place
- Node collisions happen, priority rules might have to be applied
- These details don't affect the overall process

Two simulations t=3, 150x150 grid 10,000 blue and 10,000 red agents





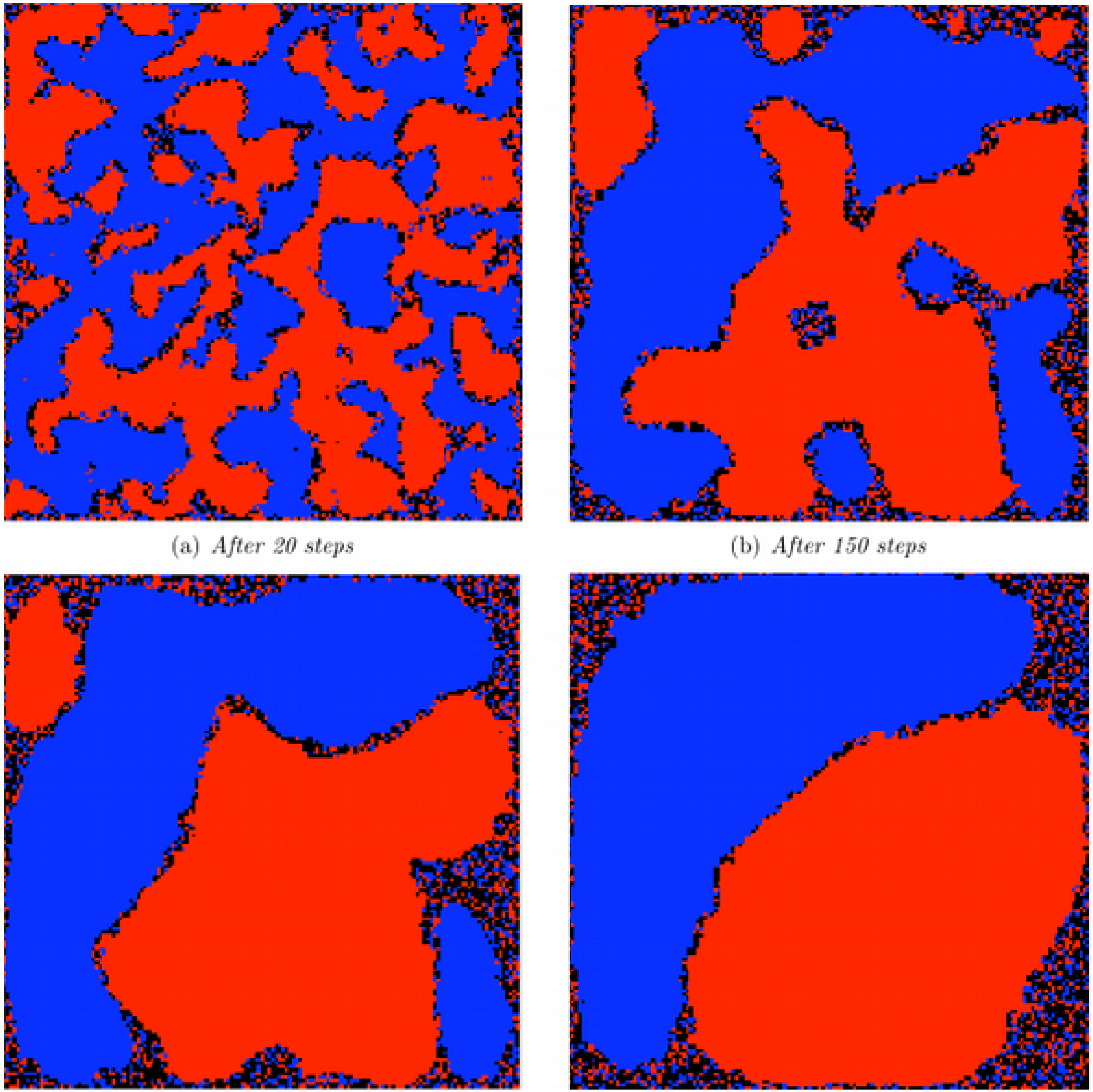
Large vs small clusters

- **In theory** agents could just form many small clusters, so one could have neighborhoods that are integrated, with small sub-groups inside
- However, **in practice** they tend to join large clusters, hence neighborhoods become segregated completely

Simulations with t=4

In general, this shows that something **fixed** (race) …

… can determine something **mutable** (location, and hence connections in the lattice)



(a) *After 20 steps*

(b) *After 150 steps*

(c) *After 350 steps*

(d) *After 800 steps*