Global Diffusion via Cascading Invitations: Structure, Growth, and Homophily

Ashton Anderson
Stanford University

Daniel Huttenlocher
Cornell University

Jon Kleinberg
Cornell University

Jure Leskovec
Stanford University

Mitul Tiwari
LinkedIn Corporation

WWW 2015

Author Profile

Ashton Anderson is a Post-doctoral Researcher at ***Microsoft Research***. Before that he was a PhD student advised by Jure Leskovec at ***Stanford University***. He will start as an assistant professor in Computer Science at the ***University of Toronto*** in Fall 2017.

- He is broadly interested in research that ***bridges the gap*** between ***computer science*** and the ***social sciences***.

- Publications: ***WWW 2015***, ***WWW 2014*** (Best Paper runner-up), ***WWW 2013***, ***KDD 2012***, ***ACL 2012***, ***WSDM 2012***, ***Sociological Science***, ***Management Science*** and so on…

Author Profile

Daniel Huttenlocher is the founding Dean and Vice Provost of ***Cornell Tech***.

- He has a mix of ***academic and industry background***, having worked at the ***Xerox Palo Alto Research Center (PARC)*** and served as CTO of ***Intelligent Markets***, as well as being a faculty member at ***Cornell*** for over two decades.

- He received his bachelor's degree from the **University of Michigan** and both his Master's and Doctorate degree from ***Massachusetts Institute of Technology (MIT)***.

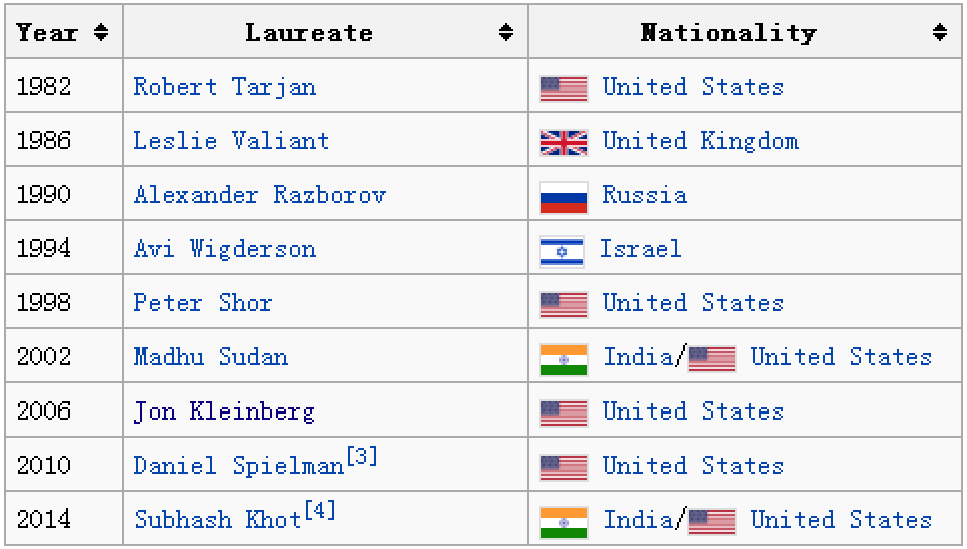
- His research on the ***web and large-scale social networks*** is focused on developing models and measures that allow us to better study and understand ***how people interact with one another***, particularly in computer-mediated environments.

Author Profile

Jon Kleinberg is a professor at Cornell University. His research focuses on issues at the interface of ***networks and information***, with an emphasis on the ***social and information networks*** that underpin the Web and other online media.

- He is a recipient of the ***Nevanlinna Prize*** by the International Mathematical Union.

- HITS algorithm.



Year	Laureate	Nationality
1982	Robert Tarjan	United States
1986	Leslie Valiant	United Kingdom
1990	Alexander Razborov	Russia
1994	Avi Wigderson	Israel
1998	Peter Shor	United States
2002	Madhu Sudan	India/ United States
2006	Jon Kleinberg	United States
2010	Daniel Spielman[3]	United States
2014	Subhash Khot[4]	India/ United States

Author Profile

Jure Leskovec is an assistant professor of computer science at ***Stanford University***. His research focuses on ***mining and modeling large social and information networks***, their ***evolution***, and ***diffusion of informa***tion and ***influence*** over them.

- He is also Chief Scientists at ***Pinterest***, where he is focusing on machine learning problems. I co-founded of a machine learning startup ***Kosei***, which was acquired by Pinterest.
- Publications: ***KDD 2015/2014/2013, AAAI 2015/2014/2013, WWW 2015/2014/2013, SIGMOD 2015, TKDD 2014, ICML 2013, ACL 2013*** and so on.

Author Profile

Mitul Tiwari is a computer scientist and a software engineer based in Silicon Valley. Currently, he is part of Search, Network, and Analytics Group at ***LinkedIn*** as a Staff Engineer (Research and Software Development).

- Previously, he worked at ***Kosmix*** as a Member of Technical Staff from October 2007 to January 2011.
- Publications: ***WWW 2015***, ***KDD 2014, VLDB 2013, WWW 2013, CIKM 2012, SIGIR 2012, IPDPS 2007*** and so on.

WWW 2015

Focus on:

- Behavioral Analysis and Personalization
- Crowdsourcing Systems and ***Social Media***
- Content Analysis
- Internet Economics and Monetization
- Pervasive Web and Mobility
- Security and Privacy
- Semantic Web
- ***Social Networks and Graph Analysis***
- Web Infrastructure: Datacenters, Content Delivery Networks, and Cloud Computing
- ***Web Mining***
- Web Search Systems and Applications

- Acceptance rate ***14.1%*** = 131/929.
- ***1400*** participants from ***77*** different countries.

Information diffusion in social networks

- Information diffusion model
  - ✓ Based on network structure
    - ➢ Linear Threshold Model
    - ➢ Independent Cascade Model
    - ➢ Other expansion models…
  - ✓ Based on state of group
    - ➢ SI, SIS, SIR
    - ➢ Other expansion models…
- Information diffusion in different networks
  - ✓ SW, SF
  - ✓ in other topologies…
- Understanding social phenomenon
  - ✓ Social Influence
  - ✓ Structural Holes
  - ✓ Conformity
  - ✓ Homophily
  - ✓ …

{ **understand**
predict
control }

Information diffusion in social networks

- Macroscopic
  - ✓ Situation
  - ✓ Breadth
  - ✓ Depth
  - ✓ Speed
  - ✓ Outbreak
  - ✓ …
- Microcosmic
  - ✓ Link prediction
    - ➢ Based on similarity
    - ➢ Based on likelihood analysis
    - ➢ …
  - ✓ Factors
    - ➢ Structure, Temporal, Location, Profile
    - ➢ Content
    - ➢ Relation
    - ➢ …

{ **understand**
predict
control }

Information diffusion in social networks

- Controllability
  - ✓ Network Topology
  - ✓ Persistent Time
  - ✓ Connectivity
  - ✓ Robustness
  - ✓ …
- Control method
  - ✓ Immunization
    - ➢ Random immunization
    - ➢ Targeted immunization
    - ➢ Acquaintance immunization
    - ➢ …
  - ✓ Cascading Failure
  - ✓ Active control
  - ✓ Positive control
  - ✓ …

{ **understand**
predict
control }

Back to this paper

Global Diffusion via Cascading Invitations: Structure, Growth, and Homophily

Ashton Anderson
Stanford University

Daniel Huttenlocher
Cornell University

Jon Kleinberg
Cornell University

Jure Leskovec
Stanford University

Mitul Tiwari
LinkedIn Corporation

WWW 2015

Outline

Background

Motivation

Global Diffusion via Cascading Invitations

Conclusions

Background

- Many of the world’s most popular websites catalyze their growth through ***invitations from existing members***.

- New members can then in turn issue invitations, and so on, creating ***cascades*** of member signups that can spread on a global scale.

- Several large sites (including **Gmail**) began with a period where this type of diffusive growth was ***the exclusive path for new signups***.

- Other sites (including LinkedIn and many others) have grown through a mix of ***cascading signups*** and ***direct signups*** at the site.

Outline

Background

Motivation

Global Diffusion via Cascading Invitations

Conclusions

Motivation

- Although these diffusive invitation processes are critical to the popularity and growth of many websites, they have ***rarely been studied***, and their properties remain ***elusive***.

- It is not known:

  ✓ How viral these cascades ***structures*** are.

  ✓ How cascades ***grow*** over time.

  ✓ How diffusive growth affects the resulting distribution of member ***characteristics*** present on the site.

Highlights

- They study ***the diffusion of LinkedIn***, an online professional network comprising over ***332 million*** members.

- They analyze the ***structural patterns*** of these signup cascades, and find them to ***be qualitatively different*** from previously studied information diffusion cascades.

- They also examine ***how signup cascades grow over time***, and observe that diffusion via invitations on LinkedIn occurs over ***much longer timescales*** than are typically associated with other types of online diffusion.

- They connect the cascade structures with ***rich individual-level attribute*** data to investigate the interplay between the two.

- They use novel techniques to study the role of ***homophily*** in diffusion. They find striking ***differences*** between ***the local***, edge-wise homophily and ***the global***, cascade-level homophily.

Outline

Background

Motivation

Global Diffusion via Cascading Invitations

Conclusions

LinkedIn Signup Cascades

- There are two ways in which a user can join LinkedIn: he can either sign up directly at the site (***a cold signup***), or he can accept an invitation from an existing LinkedIn member (***a warm signup***).

- Every ***cold signup is the root*** of its own (potentially trivial) tree, every ***warm signup has exactly one parent***, and ***cycles are impossible*** because edge sources always join earlier than their destinations.

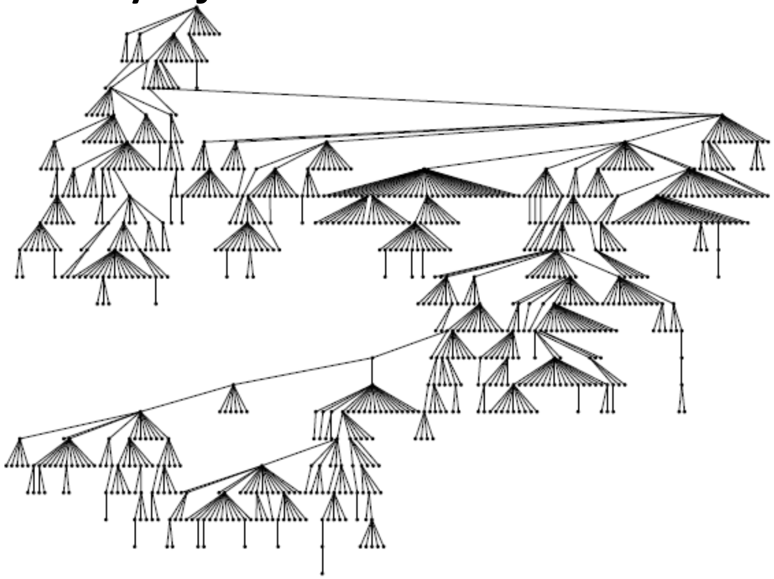


Figure 1: Example LinkedIn signup cascade.

Notes

- The LinkedIn signup forest is particularly ***well-suited*** to this paper:
  - ✓ Every member signup is ***recorded*** and ***timestamped***.
  - ✓ The diffusion of LinkedIn from one member to another is ***unambiguous***: every warm signup has a unique parent.
  - ✓ With ***over 332 million users***, LinkedIn is one of the most successful membership-based sites on the Web, and ***a large fraction of its members registered via invitations***.

Quantifying Virality of LinkedIn Cascades

- They restrict their attention to nodes at depth ***at least 1***.
- They observe that a substantial fraction of warm signups occur far from the root: for example, ***30%*** of warm signups on LinkedIn occur ***at depth 5 or greater***.
- Comparison: ***less than 1%*** of adoptions in the distributions from this earlier work are ***at depth 5 or greater***.

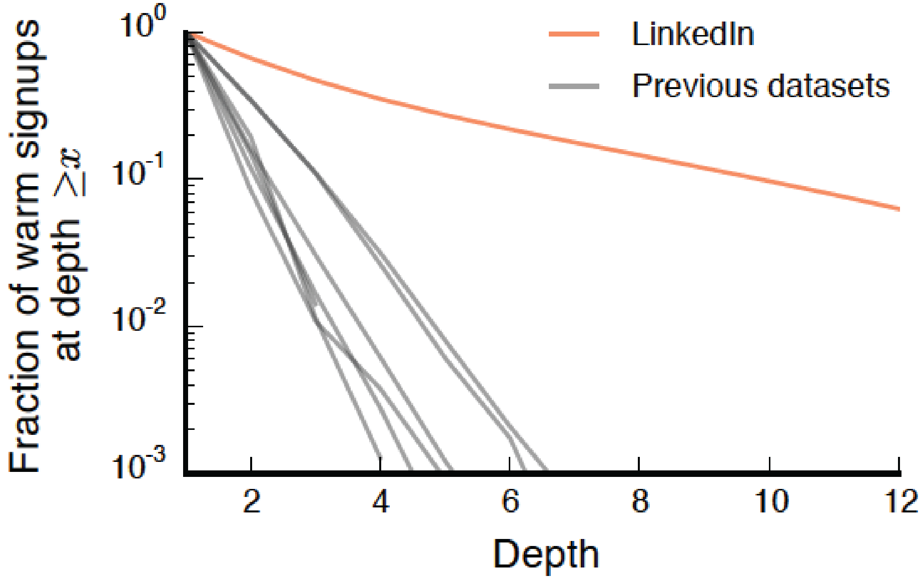
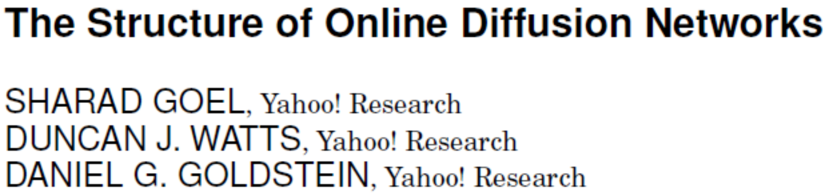
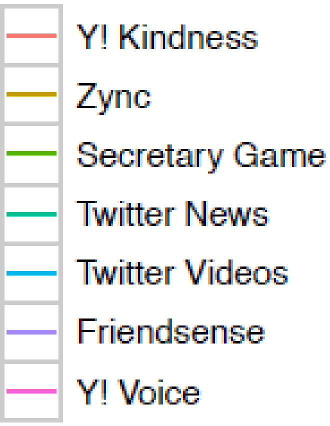


Figure 2: Distribution over adoption depth, excluding root nodes. LinkedIn adoptions occur much further from the root.



The Structure of Online Diffusion Networks

SHARAD GOEL, Yahoo! Research
DUNCAN J. WATTS, Yahoo! Research
DANIEL G. GOLDSTEIN, Yahoo! Research



Quantifying Virality of LinkedIn Cascades

- ***40%*** of non-singleton members are part of cascades with ***over 100 nodes***, whereas the same ratio is at most ***around 20%*** in the previous datasets.
- ***10%*** of non-singleton members reside in cascades with at least ***10,000*** members, whereas the largest cascades in many previous studies only have ***around few hundred*** nodes.
- On LinkedIn ***36%*** of non-singleton members reside in trees with ***maximum depth 6 or greater***, whereas the fraction in previous datasets varies ***between 0.1% and 6%.***

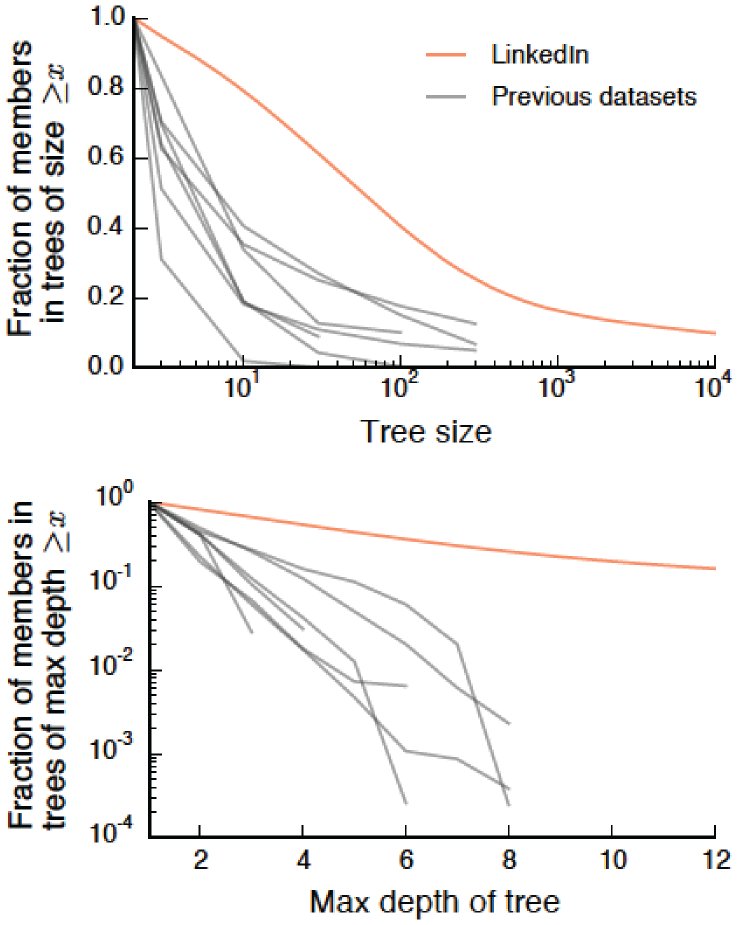


Figure 3: Fraction of non-singleton members in trees of specific size and depth. A greater portion of the LinkedIn signup forest is concentrated in large and deep cascades compared to previously studied diffusion datasets.

Structural Virality of Signup Cascades

- The structural virality measure, called the ***Wiener index***, is equal to the average path distance between two nodes in the tree.
- ***Low structural virality*** corresponds to ***broadcast-dominated*** diffusion, whereas ***high structural virality*** corresponds to ***multi-step transmission***.

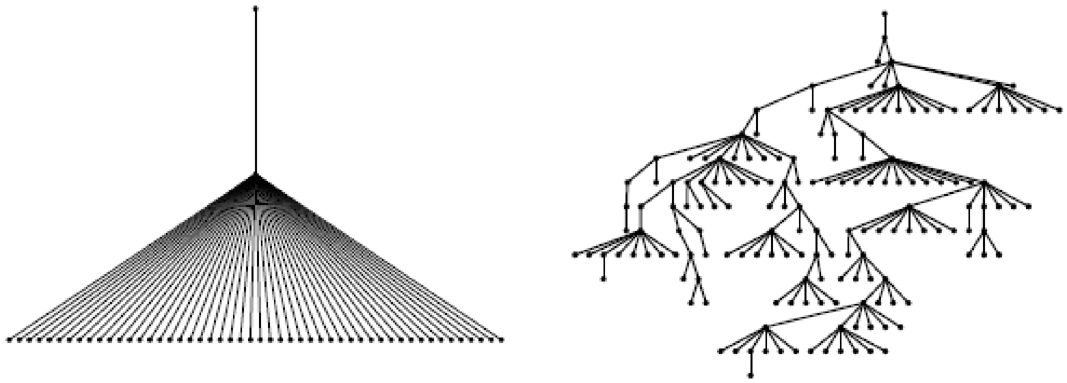


Figure 4: Two LinkedIn signup cascades, one with (left) low structural virality (Wiener index = 1.99), and one with (right) high structural virality (Wiener index = 9.5).

The Structural Virality of Online Diffusion

Sharad Goel, Ashton Anderson
Stanford University, Stanford, California, 94305 {scgoel@stanford.edu, ashton@cs.stanford.edu}
Jake Hofman, Duncan J. Watts
Microsoft Research, New York, New York 10016 {jmh@microsoft.com, duncan@microsoft.com}

Structural Virality of Signup Cascades

- A central finding in earlier analysis is that, for cascades across all major domains on Twitter, the ***correlation*** between ***structural virality*** and ***size*** is surprisingly ***low***, ***ranging between 0 and 0.2***.
- In contrast, for LinkedIn signup cascades the ***correlation*** is a strikingly ***high 0.72***.
- There are very ***few examples*** of a member "***broadcasting***" LinkedIn to hundreds or thousands of others, whereas on sites like Twitter this type of mass adoption from a single influential member is far more ***prevalent***.

Pearson Correlation Coefficient

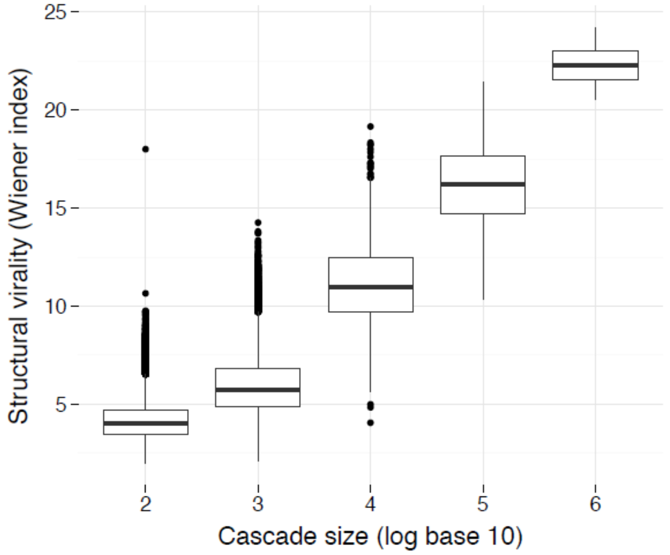
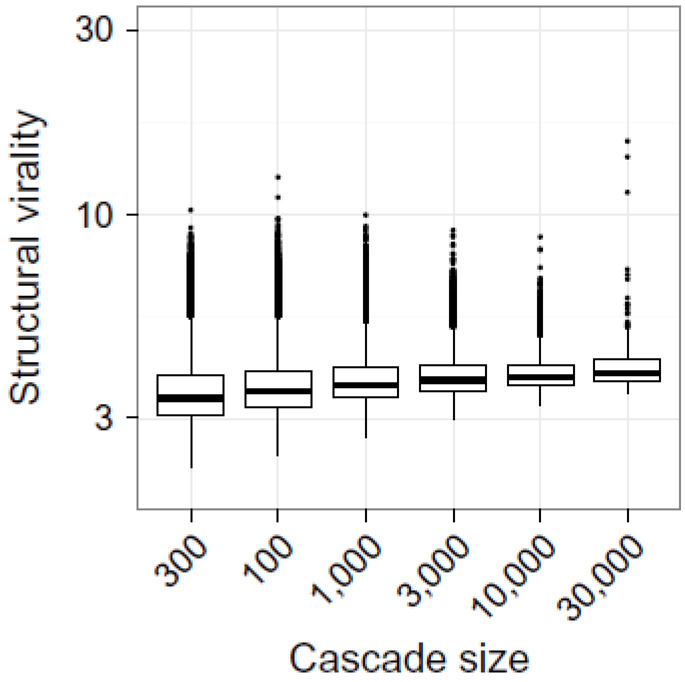


Figure 5: Structural virality as a function of cascade size (log base 10). The correlation is remarkably high, in contrast with previous findings on information-sharing cascades.



Local and Global Homophily

- What is the ***interplay*** between ***the diffusion structures*** we observe and ***the attributes*** of people involved in the diffusion process?

- ***Homophily***, the tendency of people to associate with others like themselves, it is natural to expect that much of LinkedIn's diffusion is homophily-driven.

- But, is the ***level of homophily*** between inviters and invitees, when propagated over entire cascade trees, sufficient to account for the ***global level of homogene***ity that we see in the trees as a ***whole***?

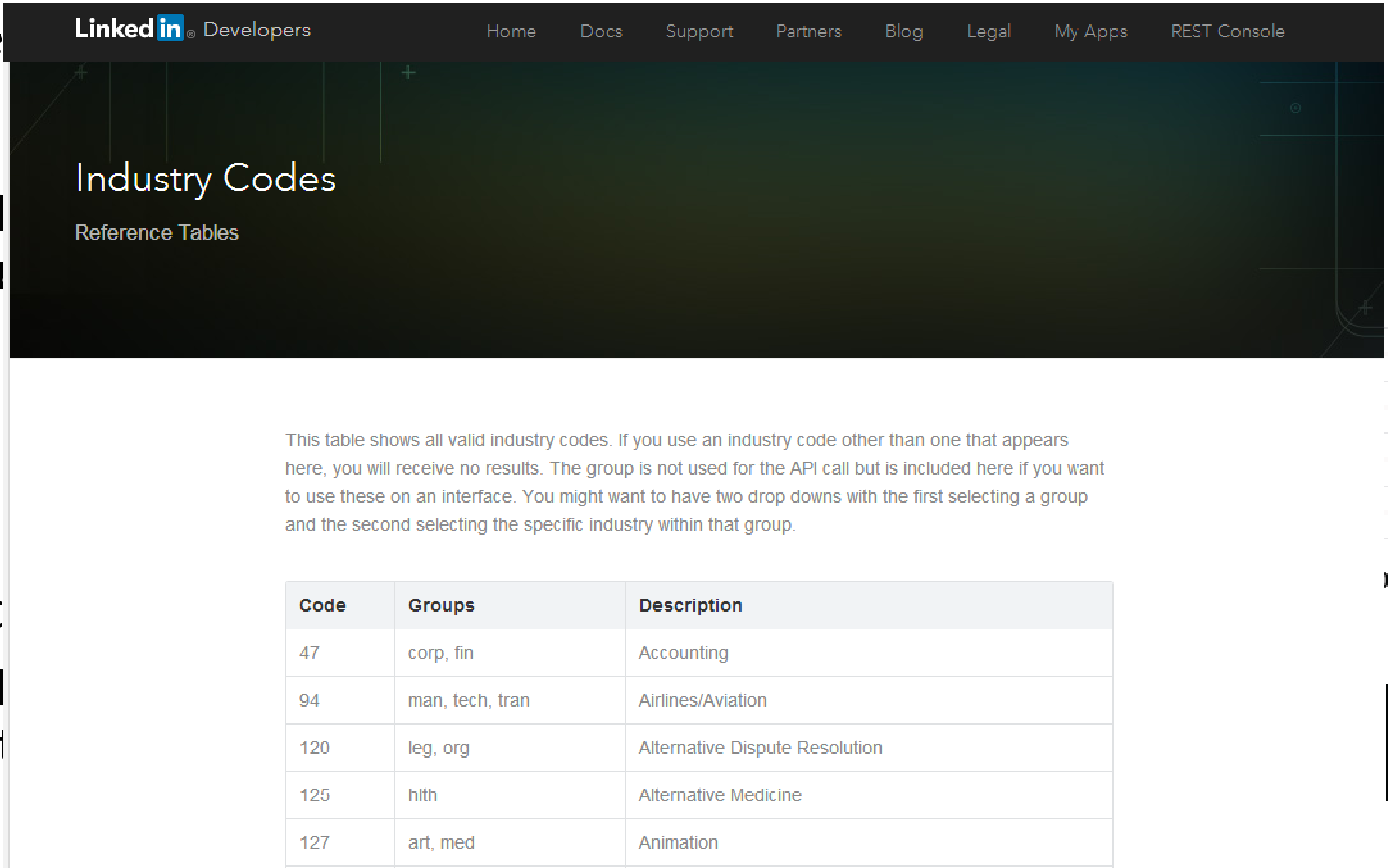
- LinkedIn is ***an ideal domain*** to study this question for two reasons:
  - ✓ They have observed a high prevalence of ***multi-step diffusion***;
  - ✓ There is a wealth of individual-level ***attribute data*** available, such as ***country of residence, geographic sub-region, professional industry of employment, age, job type, job seniority level, and others***.

Homophily in LinkedIn Signups

- Edge homophily:
  - ✓ They check this straightforwardly by computing, for every pair of attribute values A_1 and A_2, the ***conditional probability $P(A_2|A_1)$*** that a warm signup has attribute value A_2 given that their inviter has attribute value A_1.
  - ✓ This is simply equal to the empirical fraction $N(A_1 \rightarrow A_2) / N(A_1)$.
  - ✓ ***P(Brazil, Brazil)*** and ***P(India, India)*** are both ***greater than 0.80***.
- Cascade homophily:
  - ✓ They fulfill these desiderata by adopting ***the population diversity measure*** used in sociology.
  - ✓ The ***within-similarity $W_A(T)$*** of a group ***T*** on a particular attribute ***A*** is the probability that two randomly selected members match on attribute ***A***.
  - ✓ The ***between-similarity $B_A(T_1, T_2)$*** of two groups ***T_1*** and ***T_2*** is the probability that a randomly selected member from the first population and a randomly selected member from the second population match on attribute ***A***.

Homophily in LinkedIn Signups

- If there were no cascade homophily on *A* at all, then the within-tree and betwe



Industry Codes

Reference Tables

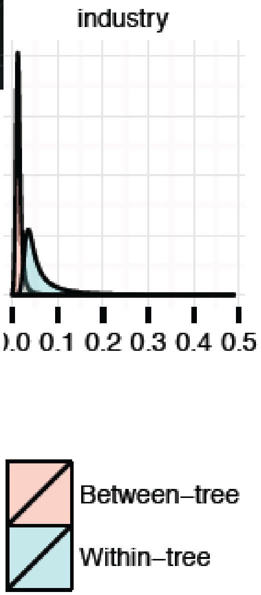
This table shows all valid industry codes. If you use an industry code other than one that appears here, you will receive no results. The group is not used for the API call but is included here if you want to use these on an interface. You might want to have two drop downs with the first selecting a group and the second selecting the specific industry within that group.

Code	Groups	Description
47	corp, fin	Accounting
94	man, tech, tran	Airlines/Aviation
120	leg, org	Alternative Dispute Resolution
125	hlth	Alternative Medicine
127	art, med	Animation

- The homo| *attribu* trees *close* overla

- Indust homo| almost

- The ***geographic attributes*** display an intriguing pattern: their within-tree similarity distributions are ***bimodal***.



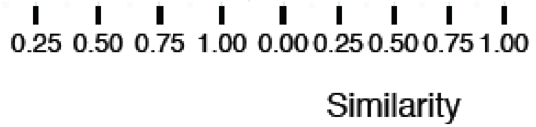


Figure 6: Within-tree and between-tree similarity on country, region, industry, engagement, and maximum job seniority.

Homophily by root country

- All five countries show ***a high degree of similarity***.

- The similarity distribution is ***unimodal*** for almost every country in our dataset. A few countries, such as France, have strong ***bimodality***.

- The overall bimodality is related to the ***diversity*** in country size, with the resulting cascade similarity ***depending on where the cascade is rooted***.

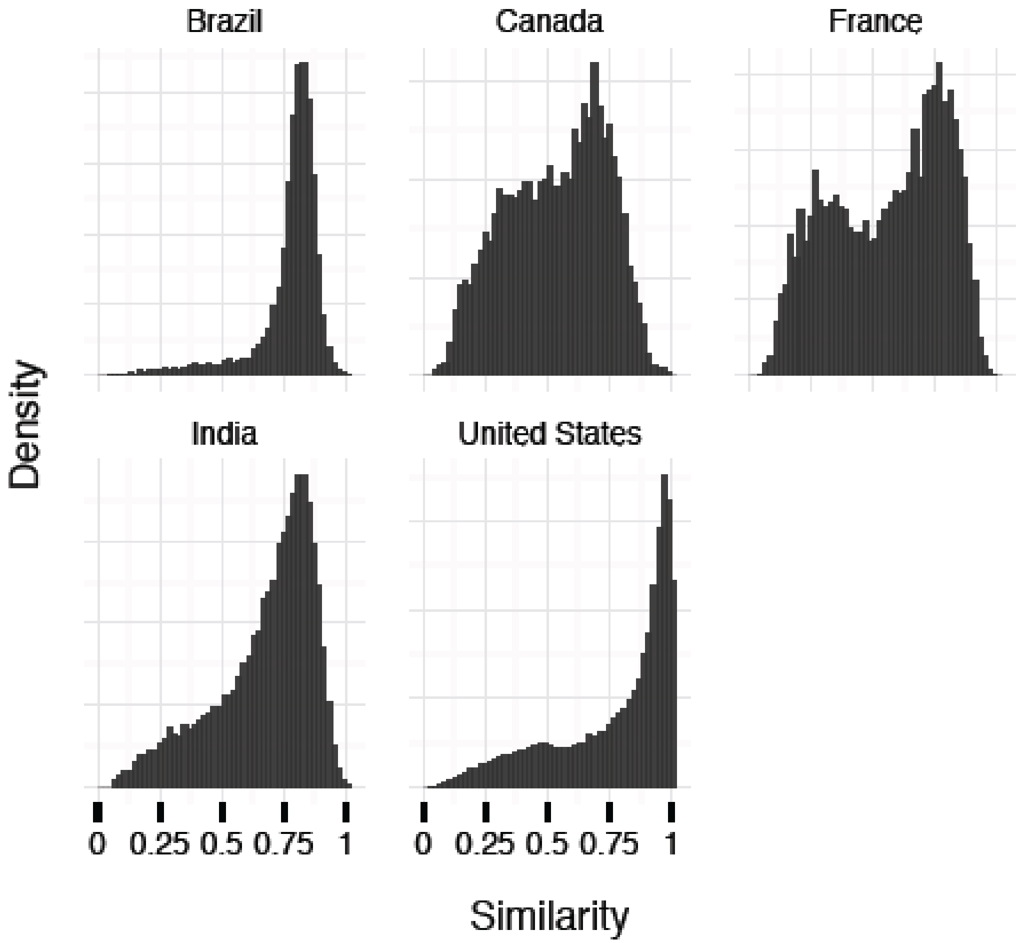


Figure 7: Within-tree similarity for trees rooted in Brazil, Canada, France, India, and the US.

Levels of Homophily

- It is unclear whether the homophily effects present in the ***signup cascades*** are different from the homophily present at a ***local level***.

- Modeling ***edge homophily***: ***First-order effects***.
  - ✓ The first-order Markov chain $\boldsymbol{M_1}$ is defined with the conditional probabilities $\boldsymbol{P(A_2|A_1)}$ computed in the previous section as transition probabilities.

- Modeling ***cascade homophily***: ***Second-order effects***.
  - ✓ The Second-order Markov chain $\boldsymbol{M_2}$ is defined by a process: the conditional probability that a new member with attribute value $\boldsymbol{A_3}$ joins, given that her inviter has value $\boldsymbol{A_2}$ and her inviter's inviter has value $\boldsymbol{A_1}$, is $\boldsymbol{P(A_3|A_1,A_2)} = \boldsymbol{N(A_1 \rightarrow A_2 \rightarrow A_3)/N(A_1 \rightarrow A_2)}$.
  - ✓ Where $\boldsymbol{N(\cdot)}$ again refers to the number of signup paths connecting nodes with particular attributes.
  - ✓ If $\boldsymbol{N(A_1 \rightarrow A_2)}$ is too small, then we ignore the grandparent and use the first-order probability $\boldsymbol{P(A_3|A_2)}$.

Levels of Homophily

- The distribution of similarity across trees is ***bimodal***, just as it is in the empirical data. This implies that ***edge homophily is sufficient to explain the bimodality*** in within-cascade similarity.
- The absolute level of within-tree similarity in the Markov simulation, is significantly ***lower than*** what we observe in empirical data.

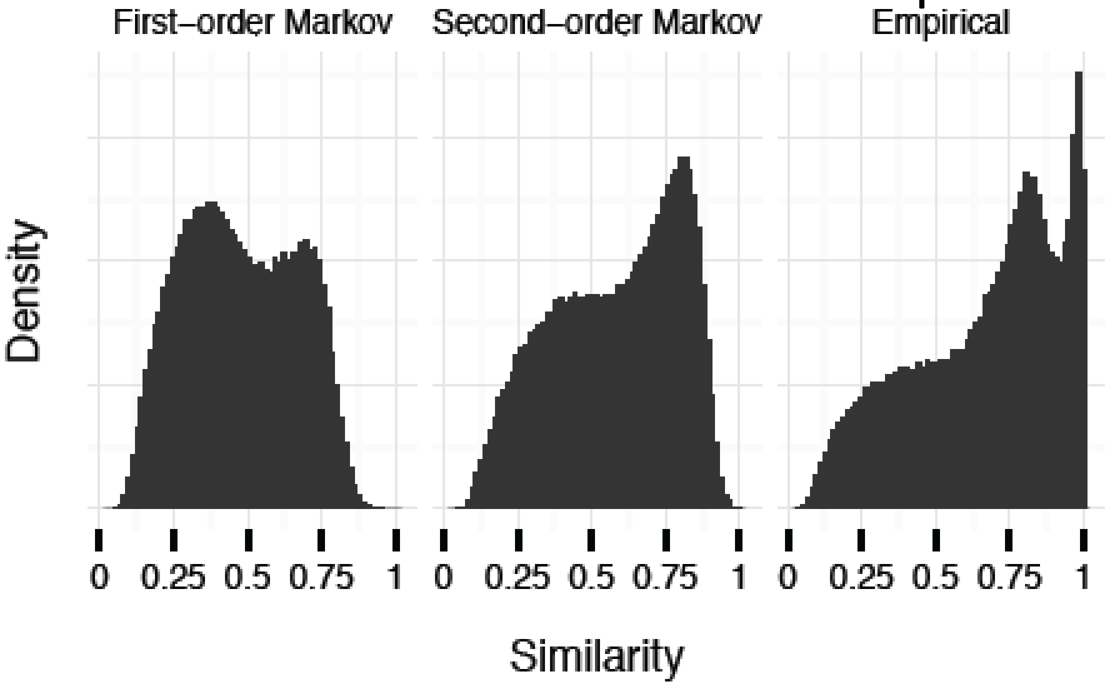


Figure 8: Within-tree similarity on real tree topologies with countries drawn from: (left) first-order Markov transitions M_1, and (middle) second-order Markov transitions M_2; (right) empirical within-tree similarity.

Guessing the Root of a Cascade

- Here they ask: how quickly does a cascade "***lose***" the attribute of the root node as the cascade grows and ***relaxes to the background distribution***?

- They consider the following concrete "***root-guessing***" question for the trees in cascade, and for the values of a particular attribute:

  - ✓ For each depth *d*, ***how often*** does the plurality attribute among members at depth *d* ***match the root's attribute***?

Guessing the Root of a Cascade

- It takes a surprisingly long time for the attributes to fully relax to the background distribution: the empirical curve only intersects the global prior at ***depth 18***.
- The first-order Markov simulation relaxes to the global prior ***much faster than*** the empirical data does.
- The second-order Markov chain ***fares significantly better***, again showing the strong higher-order homophily interactions present in signup cascades.

For example, if the parent is someone who moved from ***India to the US*** and simply ***lists the US*** as their country, then there may be information in the fact that ***the grandparent lists India*** as their country.

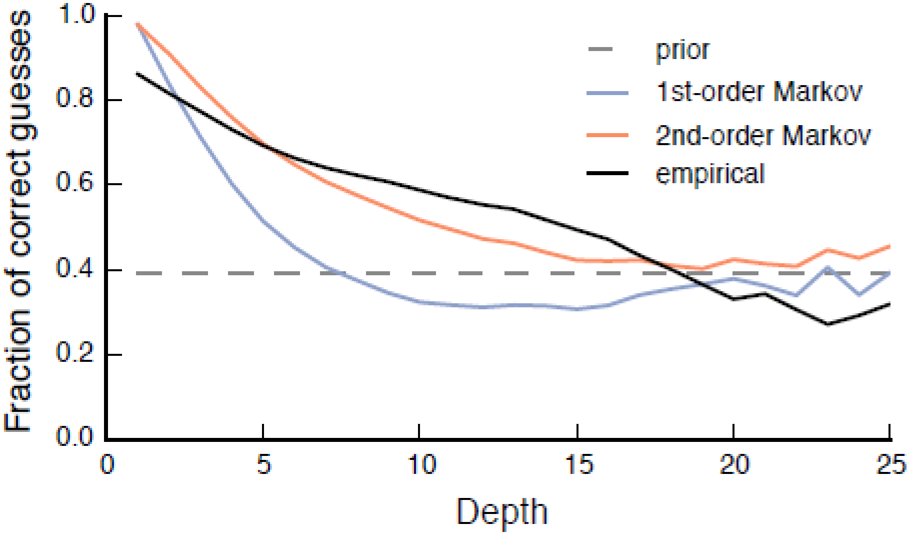


Figure 9: Fraction of time plurality attribute at depth d matches root attribute in root-guessing experiment. Empirical data retains "memory" of the root longer than baselines.

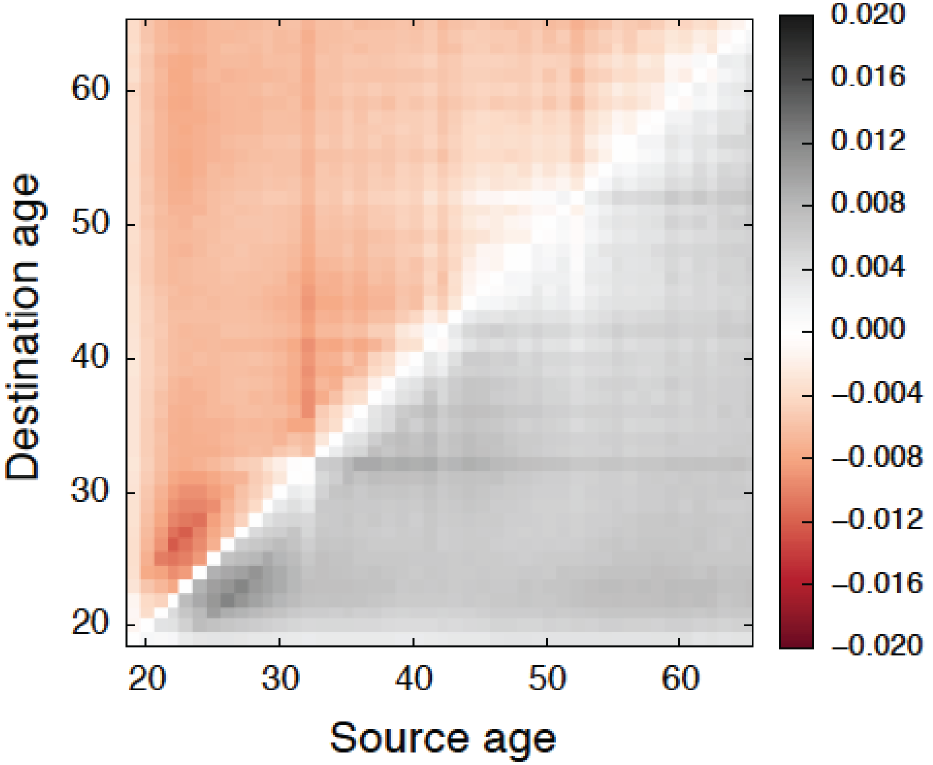
Status Gradients

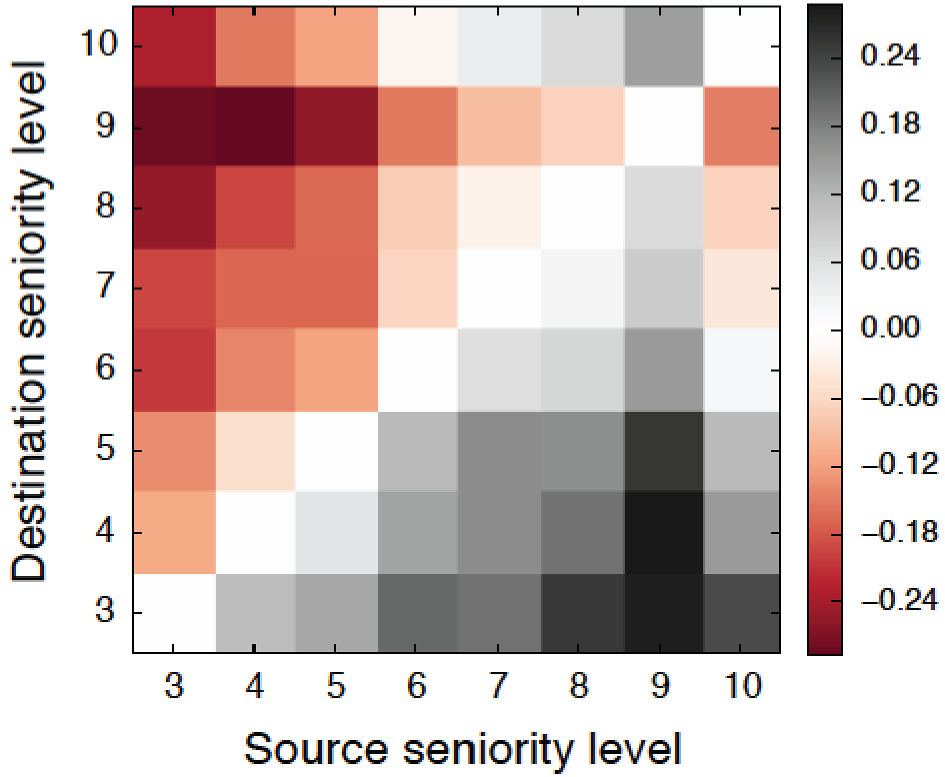
- It is possible that on attributes with ***natural orderings***, like ***age and job seniority***, signups follow a ***status gradient***, meaning people have a tendency to accept invitations from higher-status members.

- The color of the cell *(x, y)* shows how much more likely a member of type *x* is to send an accepted invitation to a member of type *y* than to receive and accept one (i.e., it is equal to ***P(y|x) – P(x|y)***.)

Status Gradients

- ***Younger*** members are more likely to accept invitations from ***older*** members than vice versa.
- They show that an even ***stronger status gradient*** exists on job seniority. (since members may have been employed in more than one job, and thus at more than one job seniority level, they define a member's seniority to be ***the highest level*** they've ever worked at)





Timescales of transmission

- A ***key characteristic*** of any diffusion process is ***how much time elapses*** between adjacent adoptions.
- They consider ***a cohort of members*** who joined LinkedIn at roughly the same time, and collect ***all signup edges (A,B)*** where *A* is a member of the cohort.
- Around ***40%*** of members who joined did so ***at least a year later*** than their inviters did. (who joined in 2006)

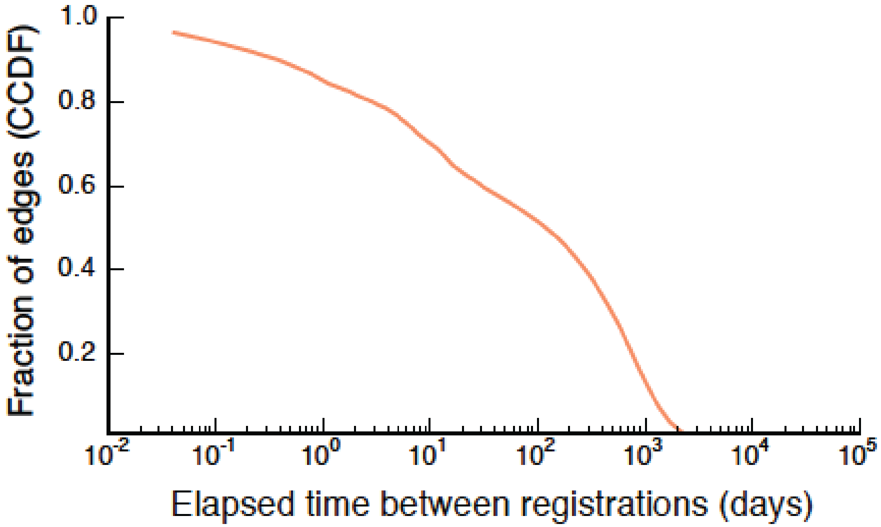


Figure 11: Complementary cumulative distribution function (CCDF) of elapsed times between inviter and invitee signup times. Adoptions are usually very separated in time.

Timescales of transmission

- Long time spans between inviter and invitee signups could be caused by two different mechanisms:
  - ✓ Members could be sending out invitations long after they register.
  - ✓ Invitees could be accepting invitations long after they receive them.
- They find that ***the former explanation*** is the case: invitations are ***sent months or years*** after members join, and invitees accept them usually ***within a few days*** after they receive them.

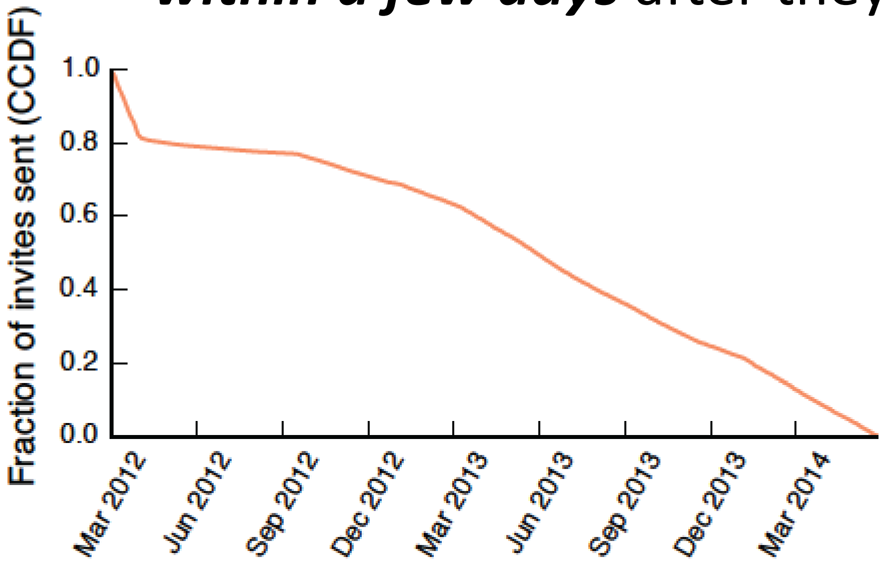


Figure 12: CCDF of the number of invites sent as a function of time for members who joined in March 2012. Most invitations are sent months after a member joins, meaning members remain "infectious" over very long periods of time.

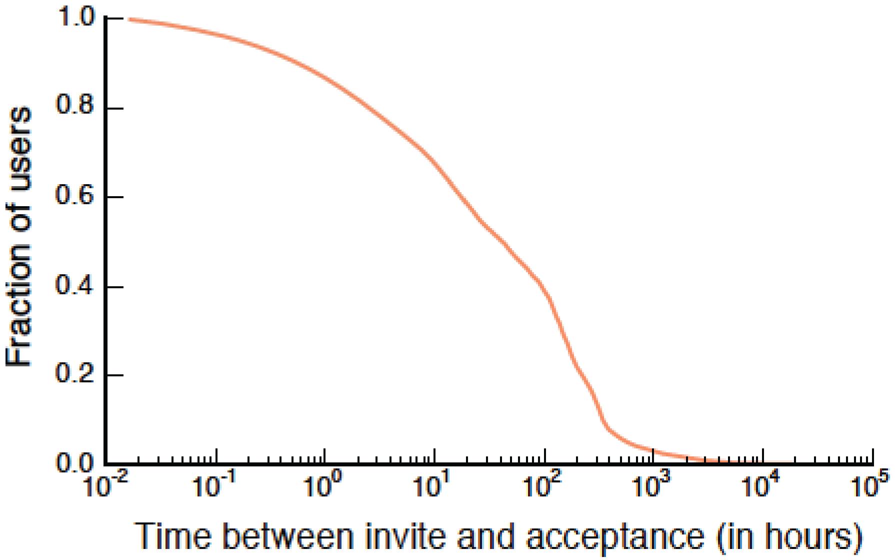
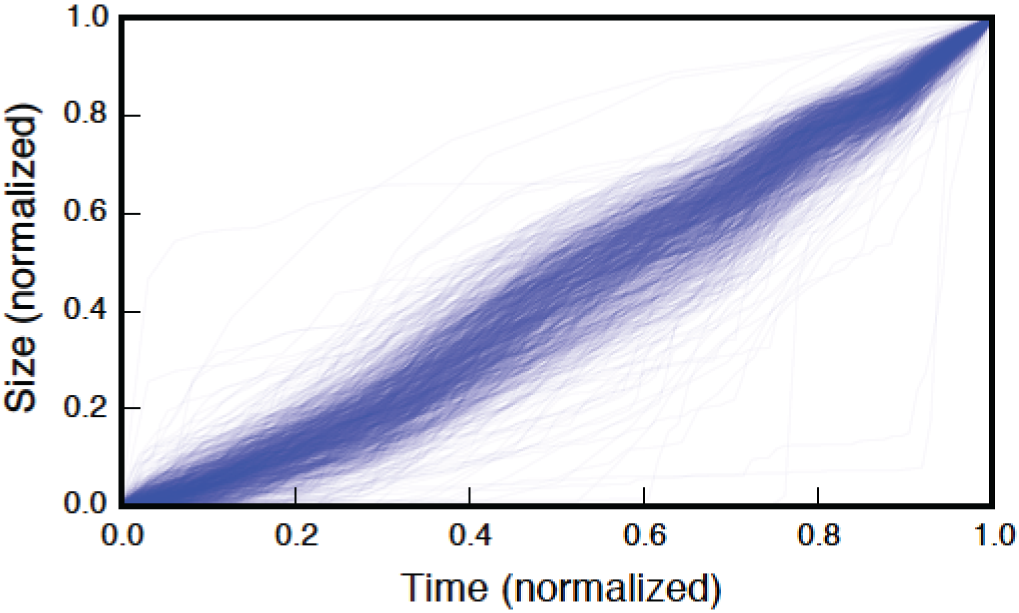


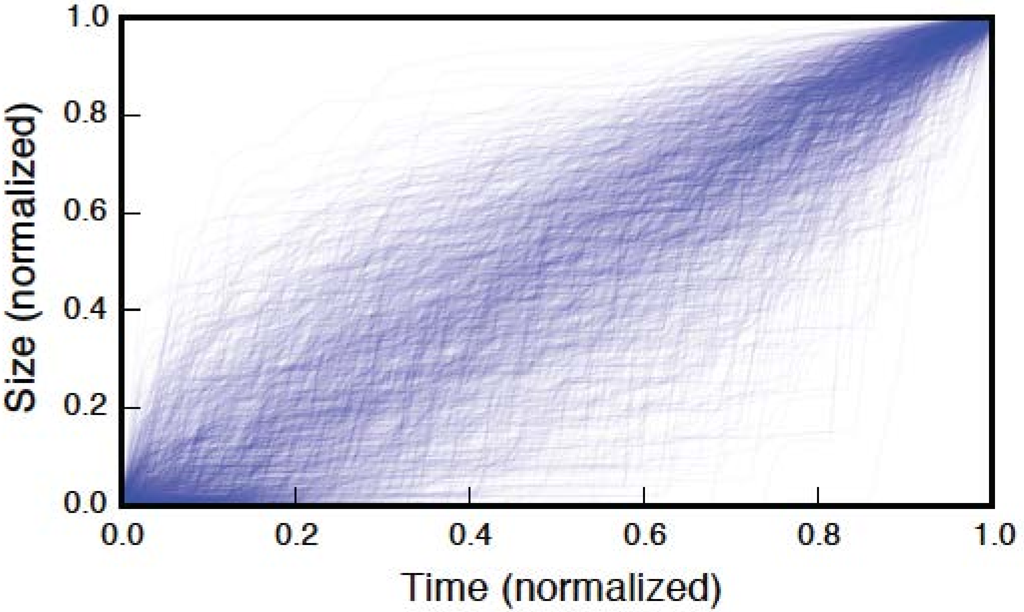
Figure 13: CCDF of the elapsed time between when an invitation was sent and when it was accepted. Invitations are accepted very quickly after they are sent.

Cascade growth trajectories

- How do LinkedIn cascades ***grow in size over time***?

- They plot the growth trajectory of the ***1,000 biggest cascades*** and ***1,000 medium-sized cascades*** on LinkedIn. For each cascade, we ***normalize*** both time and size to be between 0 and 1.
- A surprisingly robust ***linear growth pattern*** is apparent.
- LinkedIn's rapid growth is ***not*** accounted for by ***individual cascades alone***—it is the number of distinct cascades growing in ***parallel***.





Comparison with random baseline

- Arguably the simplest such baseline model is to have nodes arrive sequentially, each identified as a cold or warm signup; ***a cold signup becomes the root of a new tree, while a warm signup attaches to a parent chosen uniformly at random from existing nodes***.
- ***30 runs*** of this process.

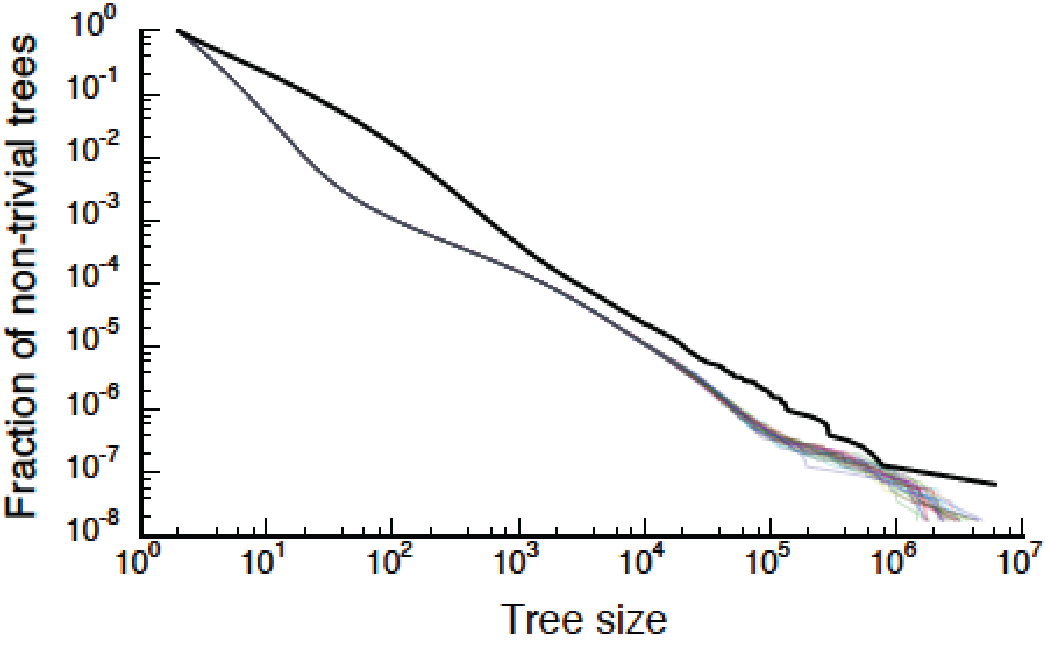


Figure 15: Distribution (CCDF) of cascade sizes on random baseline with real ordering of warm and cold signups (color) and empirical distribution of cascade sizes (black).

Outline

Background

Motivation

Global Diffusion via Cascading Invitations

Conclusions

Conclusions

- By analyzing the global spread of LinkedIn, they have been able to ***formulate and address a broad set of questions*** about ***signup cascades***—large diffusion events in which users become members of a Web site and invite friends to join as well.

- They found that the trees of signups arising from this process have characteristic structure and growth dynamics that ***look very different*** from the large information-sharing cascades that have been studied extensively in recent work.

- They also provide ***a new framework for analyzing homophily*** in these types of processes, identifying connections between the way homophily operates at multiple levels of scale.