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

Ashton Anderson
Stanford

Daniel Huttenlocher, Jon Kleinberg, Jure Leskovec, Mitul Tiwari
Cornell
Cornell
Stanford
LinkedIn

Friday, May 1, 15
growth via cascading signups

many successful websites grow by their members inviting non-members to join

e.g., Gmail, Facebook, LinkedIn, etc.
billions of accounts, huge fraction of all web traffic
questions

what’s the structure of this growth? (is it “viral”?)

how do cascades grow over time?

what types of people transmit to what types of people?
guest invitations

LinkedIn: 332M members
significant fraction are warm signups

largest product diffusion event ever analyzed
guest invitations

we construct a graph as follows:

\[ u \text{ invites } v \]

and \( v \) accepts \( u \)'s invitation
guest invitations

these invitations link together and form cascades
guest invitations

every cold signup is the root of a signup cascade

cascades are trees

all non-root nodes are warm signups
guest invitations

and

u

invites

v

accepts

u's invitation

time

Friday, May 1, 15
global diffusion via cascading invitations

1. structure
2. growth
3. homophily
cascade structure

prior work found little evidence of real multi-step, person-to-person diffusion

vast majority of “diffusion” cascades:
global diffusion via cascading invitations

1. structure
2. growth
3. homophily
cascade structure

is there evidence of “viral transmission” on LI?

one way to quantify: how many of the adopters are far from the root?
cascade structure

adoptions are much deeper on LI than in previous datasets
cascade structure

another measure: what fraction of adoptions are accounted for in large/deep cascades?
cascade structure

another measure: what fraction of adoptions are accounted for in large/deep cascades?

so much more viral transmission that we’re observing qualitatively different behavior
cascade structure

structural virality of a cascade: rigorous measure to interpolate between broadcast and viral diffusion

broadcast (low SV)  viral (high SV)
cascade structure

important question: what’s the relationship between cascade size and structural virality?

if strongly negative or positive, knowing cascade size tells you mechanism by which it grew

if close to 0, cascades grow in structurally different ways
cascade structure

prior work: Twitter information cascades

correlations range from 0.0 to 0.2
cascade structure

our work: LinkedIn signup cascades

strikingly high correlation: 0.72!
cascade structure

LinkedIn signup cascades are qualitatively different than previously studied online diffusion datasets

direct evidence of a large-scale, multi-step diffusion process
...in contrast with previous work
global diffusion via cascading invitations

1. structure
2. growth
3. homophily
growth dynamics

information cascades grow and flame out very quickly (think news, etc.)

what timescales do LI cascades operate over?
growth dynamics

time gap between inviter, invitee signups

months and years, not hours!
growth dynamics

invites sent later

invites accepted quickly

LI cascades are extremely persistent
growth dynamics

information cascades grow quickly then stagnate

LI cascades are much more persistent: what is the growth trajectory of a LI cascade?
growth dynamics

tree growth over time for 1K biggest trees surprisingly linear!
growth dynamics

LI signup cascades accruing members at a steady, persistent, constant rate

not the “burn through the network” picture of information diffusion
global diffusion via cascading invitations

1. structure
2. growth
3. homophily
extremely rich user-level data: we can now see how diffusion relates to underlying node attributes

*homophily*: the tendency for people to associate with others like themselves ("birds of a feather flock together")
homophily

we consider all cascades with \( \geq 100 \) nodes
\( (n > 100K \text{ of them}) \)

every cascade defines a set of members

look at distributions of attributes in
individual cascades
homophily

**within-similarity**: probability that two randomly chosen nodes match on an attribute

**between-similarity**: probability that a randomly drawn node from group 1 matches on an attribute with a randomly drawn node from group 2

the difference between the two is a measure of homophily
homophily

- Country
- Region
- Industry

Engagement
Maxseniority

Similarity

Between-tree
Within-tree
homophily

extreme homophily on geography

significant homophily on industry

minimal homophily on engagement, max seniority level, and age
homophily

Specifies that individuals tend to form relationships with others who are similar to them. The prevalence of homophily is illustrated through a density graph comparing different countries, showing that adoption rates are much deeper on LI than in previous datasets.
homophily

clearly, there is strong homophily on country

but does this *cascade* homophily follow from the obvious *edge* homophily?
homophily

model edge homophily with a first-order Markov chain
model edge homophily with a first-order Markov chain

empirically derived transition matrix:

<table>
<thead>
<tr>
<th></th>
<th>BR</th>
<th>CA</th>
<th>FR</th>
<th>IN</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>0.85</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.11</td>
</tr>
<tr>
<td>CA</td>
<td>0.03</td>
<td>0.60</td>
<td>0.06</td>
<td>0.06</td>
<td>0.25</td>
</tr>
<tr>
<td>FR</td>
<td>0.02</td>
<td>0.10</td>
<td>0.65</td>
<td>0.03</td>
<td>0.20</td>
</tr>
<tr>
<td>IN</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.82</td>
<td>0.12</td>
</tr>
<tr>
<td>US</td>
<td>0.05</td>
<td>0.02</td>
<td>0.01</td>
<td>0.05</td>
<td>0.87</td>
</tr>
</tbody>
</table>
homophily

model edge homophily with a first-order Markov chain

<table>
<thead>
<tr>
<th></th>
<th>BR</th>
<th>CA</th>
<th>FR</th>
<th>IN</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>0.85</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.11</td>
</tr>
<tr>
<td>CA</td>
<td>0.03</td>
<td>0.60</td>
<td>0.06</td>
<td>0.06</td>
<td>0.25</td>
</tr>
<tr>
<td>FR</td>
<td>0.02</td>
<td>0.10</td>
<td>0.65</td>
<td>0.03</td>
<td>0.20</td>
</tr>
<tr>
<td>IN</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.82</td>
<td>0.12</td>
</tr>
<tr>
<td>US</td>
<td>0.05</td>
<td>0.02</td>
<td>0.01</td>
<td>0.05</td>
<td>0.87</td>
</tr>
</tbody>
</table>

edge homophily
homophily

simulate signup diffusion with first-order Markov chain
homophily

simulate signup diffusion with first-order Markov chain
homophily

simulate signup diffusion with first-order Markov chain

<table>
<thead>
<tr>
<th></th>
<th>BR</th>
<th>CA</th>
<th>FR</th>
<th>IN</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>0.85</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.11</td>
</tr>
<tr>
<td>CA</td>
<td>0.03</td>
<td>0.60</td>
<td>0.06</td>
<td>0.06</td>
<td>0.25</td>
</tr>
<tr>
<td>FR</td>
<td>0.02</td>
<td>0.10</td>
<td>0.65</td>
<td>0.03</td>
<td>0.20</td>
</tr>
<tr>
<td>IN</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.82</td>
<td>0.12</td>
</tr>
<tr>
<td>US</td>
<td>0.05</td>
<td>0.02</td>
<td>0.01</td>
<td>0.05</td>
<td>0.87</td>
</tr>
</tbody>
</table>
homophily

simulate signup diffusion with first-order Markov chain

<table>
<thead>
<tr>
<th></th>
<th>BR</th>
<th>CA</th>
<th>FR</th>
<th>IN</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>0.85</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.11</td>
</tr>
<tr>
<td>CA</td>
<td>0.03</td>
<td>0.60</td>
<td>0.06</td>
<td>0.06</td>
<td>0.25</td>
</tr>
<tr>
<td>FR</td>
<td>0.02</td>
<td>0.10</td>
<td>0.65</td>
<td>0.03</td>
<td>0.20</td>
</tr>
<tr>
<td>IN</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.82</td>
<td>0.12</td>
</tr>
<tr>
<td>US</td>
<td>0.05</td>
<td>0.02</td>
<td>0.01</td>
<td>0.05</td>
<td>0.87</td>
</tr>
</tbody>
</table>
homophily

simulate signup diffusion with first-order Markov chain

<table>
<thead>
<tr>
<th></th>
<th>BR</th>
<th>CA</th>
<th>FR</th>
<th>IN</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>0.85</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.11</td>
</tr>
<tr>
<td>CA</td>
<td>0.03</td>
<td>0.60</td>
<td>0.06</td>
<td>0.06</td>
<td>0.25</td>
</tr>
<tr>
<td>FR</td>
<td>0.02</td>
<td>0.10</td>
<td>0.65</td>
<td>0.03</td>
<td>0.20</td>
</tr>
<tr>
<td>IN</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.82</td>
<td>0.12</td>
</tr>
<tr>
<td>US</td>
<td>0.05</td>
<td>0.02</td>
<td>0.01</td>
<td>0.05</td>
<td>0.87</td>
</tr>
</tbody>
</table>
homophily

simulate signup diffusion with first-order Markov chain

<table>
<thead>
<tr>
<th></th>
<th>BR</th>
<th>CA</th>
<th>FR</th>
<th>IN</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>0.85</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.11</td>
</tr>
<tr>
<td>CA</td>
<td>0.03</td>
<td>0.60</td>
<td>0.06</td>
<td>0.06</td>
<td>0.25</td>
</tr>
<tr>
<td>FR</td>
<td>0.02</td>
<td>0.10</td>
<td>0.65</td>
<td>0.03</td>
<td>0.20</td>
</tr>
<tr>
<td>IN</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.82</td>
<td>0.12</td>
</tr>
<tr>
<td>US</td>
<td>0.05</td>
<td>0.02</td>
<td>0.01</td>
<td>0.05</td>
<td>0.87</td>
</tr>
</tbody>
</table>
homophily

simulate signup diffusion with first-order Markov chain

<table>
<thead>
<tr>
<th></th>
<th>BR</th>
<th>CA</th>
<th>FR</th>
<th>IN</th>
<th>US</th>
</tr>
</thead>
<tbody>
<tr>
<td>BR</td>
<td>0.85</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.11</td>
</tr>
<tr>
<td>CA</td>
<td>0.03</td>
<td>0.60</td>
<td>0.06</td>
<td>0.06</td>
<td>0.25</td>
</tr>
<tr>
<td>FR</td>
<td>0.02</td>
<td>0.10</td>
<td>0.65</td>
<td>0.03</td>
<td>0.20</td>
</tr>
<tr>
<td>IN</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.82</td>
<td>0.12</td>
</tr>
<tr>
<td>US</td>
<td>0.05</td>
<td>0.02</td>
<td>0.01</td>
<td>0.05</td>
<td>0.87</td>
</tr>
</tbody>
</table>
homophily

simulate signup diffusion with first-order Markov chain
homophily

keep all cascade structures the same

run this first-order Markov chain process to generate simulated attribute distributions

compute within-similarity as before

if distribution over similarities is similar, then cascade homophily follows from edge homophily
Markov-generated similarities much lower than observed values!
homophily

this reveals a deep fact:

*LI signup cascades are not arbitrary sets of members*

that there is cascade homophily above and beyond the already-high edge homophily means that there is **higher-order structure** in the cascades
homophily

repeat the same experiment with second-order Markov chain

instead of considering just the parent, consider grandparent and parent
“second-order effects” very large here
homophily

how long-range is the dependence?

root-guessing experiment borrowed from genetics

given node attributes at depth d, does plurality attribute match root attribute?
Homophily

adoptions are much deeper on LI than in previous datasets
homophily

adoptions are much deeper on LI than in previous datasets
homophily

Adoptions are much deeper on LI than in previous datasets.

US
US
US
US
BR
US
US
BR
IN
BR
BR
homophily

run this experiment on:

- real attributes
- first-order Markov generated attributes
- second-order Markov generated attributes
homophily

Fraction of correct guesses

Depth
homophily

genetic processes are first-order by definition

higher-order dependencies in our setting is thus analogous to phenotypes, not genotypes

a member profile is like a social phenotype

what would a social genotype look like?
conclusion

LI cascades much more structurally viral than previously studied diffusion datasets

they grow persistently over time

significant homophily patterns at cascade level, meaning cascades are coherent sets of members
thank you!
status effects

![Diagram showing status effects over source and destination ages. The x-axis represents the source age, ranging from 20 to 60, and the y-axis represents the destination age, also ranging from 20 to 60. The color gradient indicates the strength of the status effects, with darker colors representing stronger effects.]