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

Ashton Anderson
Stanford University
ashton@cs.stanford.edu

Daniel Huttenlocher
Cornell University
dph@cs.cornell.edu

Jon Kleinberg
Cornell University
kleinber@cs.cornell.edu

Jure Leskovec
Stanford University
jure@cs.stanford.edu

Mitul Tiwari
LinkedIn Corporation
mtiwari@linkedin.com

ABSTRACT

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. 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. For instance, it is not known how viral these cascades structures are, how cascades grow over time, or how diffusive growth affects the resulting distribution of member characteristics present on the site.

In this paper, we study the diffusion of LinkedIn, an online professional network comprising over 332 million members, a large fraction of whom joined the site as part of a signup cascade. First we analyze the structural patterns of these signup cascades, and find them to be qualitatively different from previously studied information diffusion cascades. We 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. Finally, we connect the cascade structures with rich individual-level attribute data to investigate the interplay between the two. Using novel techniques to study the role of homophily in diffusion, we find striking differences between the local, edge-wise homophily and the global, cascade-level homophily we observe in our data, suggesting that signup cascades form surprisingly coherent groups of members.

Categories and Subject Descriptors: H.2.8 [**Database management**]: Database applications—*Data mining*.
Keywords: cascades; product diffusion; social networks.

1. INTRODUCTION

One of the central dynamics on the Web is the tremendous growth of new sites and services that expand from small sets of early adopters to huge user populations. There are several mechanisms through which this can occur, and here we consider one that has been crucial for a number of the world's largest websites: attracting participants through a cascade of signups, where current users invite others to join the site. In this process, users have the option of issuing invitations to people not yet on the site. When these invitations are accepted, the resulting new users can issue further invitations, and the process continues, potentially diffusing through multiple levels. 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.

Since many of the most popular sites acquire their large audiences in part through invitations mechanisms like this, Web users experience the effects of these diffusive growth processes on a daily basis. Despite this, however, we have very little understanding of what these processes look like at a global level, nor what consequences they have for the makeup of a site's user population.

In this paper we address questions about cascading adoption processes using a dataset that represents the complete diffusion of LinkedIn, an online professional network that contains over 332 million users. To our knowledge, this is the largest structural analysis of a cascading adoption event ever undertaken. In addition its global scale, a significant fraction of LinkedIn's growth came through invitations, making it an ideal domain for addressing core questions about successful cascading processes. We note that while the particular numerical results presented here are of course specific to LinkedIn, they clearly illustrate certain patterns that are markedly different from what has been found in earlier diffusion studies, and they point to properties that may be characteristic of other large product diffusions as well.

The Present Work: Structure and Growth of Cascading Signups. We first analyze the underlying structure and growth dynamics of the LinkedIn cascades. The cascading signups can be naturally organized into a collection of trees: each time a user signs up directly, without an invitation, she forms the root of a new tree, and every user B who accepts an invitation from a user A joins the tree containing A, as the child of A.

We contrast these trees with the kinds of trees that arise in *information-sharing cascades*, in which users pass small information units (*e.g.*, links, photos, or other memes) to their friends. Like our cascading signups here, information-sharing cascades are also a form of online diffusion, but we discover that the properties of these two types of cascades are quite different. The work on information-sharing cascades has suggested that their temporal dynamics unfold

WWW 2015, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3469-3/15/05.
http://dx.doi.org/10.1145/2736277.2741672.

very quickly, that most of their adoptions occur very close to the root, and that their size and their structural virality are essentially uncorrelated [8, 14, 13].

Our analysis of diffusion patterns in LinkedIn signups provides evidence that signup cascades, in contrast, are significantly more viral than previously-studied online diffusion datasets in a number of dimensions. Additionally, we find that size and structural virality are strongly correlated, suggesting that the largest signup cascades grow almost exclusively through viral, multi-step diffusion. Moreover, cascade growth plays out over much longer periods of time: users in a cascade remain active in spreading LinkedIn to new users months or even years after their own registration. The result is an approximately *linear* rate of tree growth for the large trees, and with a growth rate over time that is remarkably uniform across different trees. The commonality across trees suggests that there is a characteristic growth pattern for these signups that seems to be truly independent of the tree in which it occurs.

The Present Work: Homophily in Cascades. In addition to examining the architecture of the cascading adoption process itself, we also connect cascade structure with attributes of the users in the trees and investigate the interplay between the two. We find that attributes such as geography and industry play a substantial role in the cascades, exhibiting a pattern of homophily in which people share characteristics with those who invited them. But our investigations strongly suggest that this is a deeper process, arising from something more than just similarity between the inviter and invitee.

In particular, our fine-grained view of the invitation process suggests that we can think about the basic issue of homophily at two fundamental levels of scale: locally, in which users invite friends with similar characteristics, and globally, in which entire trees may or may not exhibit certain levels of homogeneity. One way to approach the connection between local and global patterns of similarity is through the following question: Is the direct similarity between inviters and invitees, when propagated through the structure of an entire tree, sufficient to account for the level of user homogeneity that we see in the full trees?

We find that, surprisingly, the local similarity between inviters and invitees is in fact *not* sufficient to produce the global levels of homophily we observe. This gap between the local and global properties of the trees raises the question of what other factors are playing a role in the global patterns of similarity that we observe.

To address this question, we ask what happens if the distribution of user attributes were based on a Markov process defined on the cascade trees, with the attributes of each node in the tree arising probabilistically from the attributes of its parent node. Such models have been considered in a very different context, in work that studies genetic inheritance with mutation on evolutionary trees [12, 31, 32, 35]. A way to summarize the disconnect between the local and global properties of our cascade trees is to say that this first-order Markov process gives rise to less homogeneity among a tree's nodes than we find in the real data. However, we find that higher-order Markov models, in which a node's attributes can depend not just on its parent but on its earlier ancestors as well, produce a level of homogeneity that more closely matches the data. This suggests that LinkedIn signup cascades are coherent, in that they are comprised of users who are more uniform than pairwise similarities can account for. The alignment with this higher-order Markov models suggests new ways of thinking about how user characteristics become correlated in these types of cascading processes.

The paper is outlined as follows: we discuss our structural analyses in Section 2, then investigate local and global homophily in Section 3 and cascade growth in Section 4, and finally we summarize related work in Section 5 and conclude in Section 6.

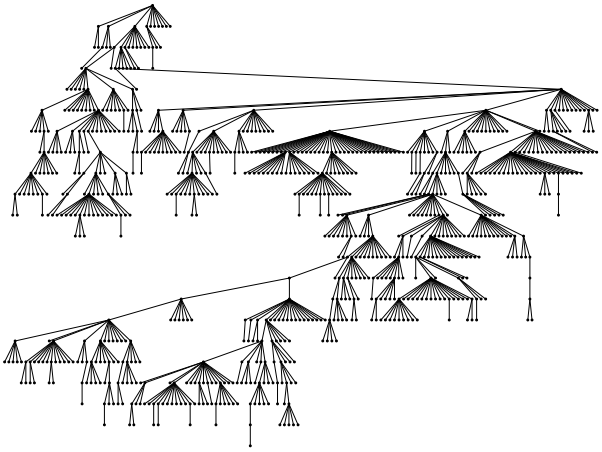


Figure 1: Example LinkedIn signup cascade.

2. STRUCTURAL CHARACTERISTICS

We begin with a structural analysis of the LinkedIn signup cascades. First we describe the invitation process and how sets of signups form cascades.

2.1 LinkedIn Signup Cascades

There are two ways in which a user B can join LinkedIn: she can either sign up directly at the site (a *cold signup*), or she can accept an invitation from an existing LinkedIn member (a *warm signup*). To count as a warm signup, it is not sufficient for B to simply have received an invitation from a member—she must also click on such an invitation and sign up through the resulting interface. The member A who sent the invitation that B used to register is then recorded, and we consider this an accepted invitation $A \rightarrow B$. At most one edge is created per signup.

We construct an accepted-invitation graph F where every member of LinkedIn is a node, and there is an edge between A and B if B joined LinkedIn by accepting an invitation from A. Note that F is a forest (a set of trees): 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. Studying the structural properties of this signup forest is the focus of this section (see Fig. 1 for an example cascade).

Before examining the data and discussing our results, we note that the LinkedIn signup forest is particularly well-suited to our goals. First, as is typical of data derived from an online platform, it is very fine-grained: every member signup is recorded and time-stamped, and the identity of the inviter (if any) is known. This allows us to perfectly reconstruct the massive LinkedIn registration process as it unfolded over time. Second, the diffusion of LinkedIn from one member to another is unambiguous: every warm signup has a unique parent. People can receive multiple invitations to join LinkedIn, but they can only use one of them to sign up—therefore, there is no ambiguity in which member is the parent. It is certainly true that there will often be other factors in an individual's decision to join beyond just the accepted invitation, including receiving other invitations, but it is also the case that the edges in our graph correspond to precisely the invitations that were actually accepted. Third, with over 332 million users, LinkedIn is one of the most successful membership-based sites on the Web, and a large fraction of

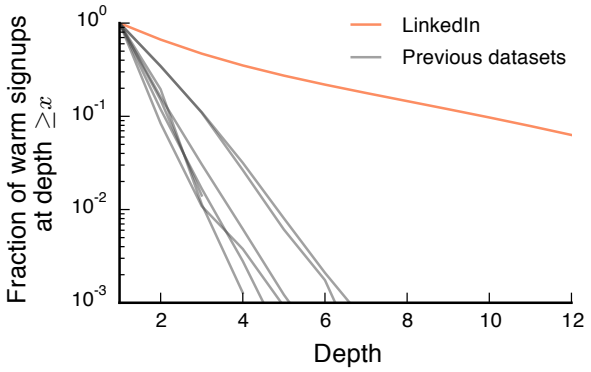


Figure 2: Distribution over adoption depth, excluding root nodes. LinkedIn adoptions occur much further from the root.

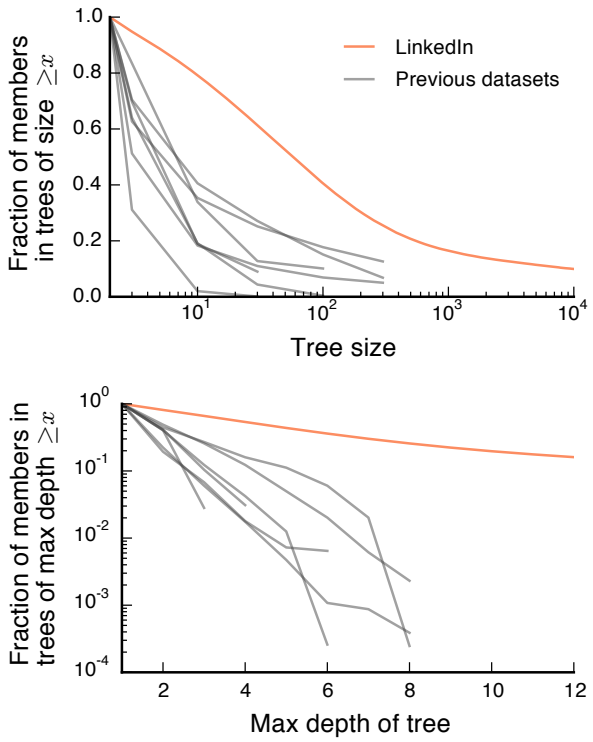


Figure 3: Fraction of non-singleton members in trees of specific size and depth. A greater portion of the LinkedIn signup forest is concentrated in large and deep cascades compared to previously studied diffusion datasets.

its members registered via invitations, which makes it an ideal subject for a case study investigating the structure of a massive online product diffusion event.

Our analyses demonstrate the contrast between the structural properties of the cascades studied in the past [14, 22, 26] and the structure of the signup cascades from a large popular website as we study here. We compare structural properties of the diffusion of LinkedIn with the diffusion of news stories, videos, online applications, and services reported in [14].

A recurring theme in our analysis is that the viral structure of LinkedIn's signup forest isn't necessarily due to its large size. In particular, many of the measures we consider can be relatively constant for cascades of widely varying sizes; the fact that the LinkedIn forest is so different for a number of these measures thus does not follow from pure scale, but instead is indicative of the nature of its propagation.

2.2 Quantifying Virality of LinkedIn Adoption

Adoptions as a function of depth. A natural way to measure the extent to which viral propagation accounts for adoption is to examine the distribution over tree depths (number of steps from the root) where adoptions occur. Since we are interested in quantifying person-to-person transmission, or warm signups, we restrict our attention to nodes at depth at least 1; the normalized distribution is shown in Fig. 2. We 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. Comparing this distribution with those reported in the previous work of [14], the difference is quite striking; for instance, less than 1% of adoptions in the distributions from this earlier work are at depth 5 or greater. (Note that our results are qualitatively the same whether or not we include root nodes, as was done in the original analysis of [14].)

Adoptions by cascade type. Another way to gauge the virality of diffusion events is by measuring what fraction of adoptions reside in deep or large cascades.

In Fig. 3(top), we show the fraction of non-singleton members that reside in trees of increasing sizes. The differences between LinkedIn signup cascades and the previously studied cascades are again substantial. For example, 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. Furthermore, 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 [14, 22, 26]. A similarly large gap exists when we consider tree depth: 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% (Fig. 3(bottom)). As before, the comparisons with the previous datasets are qualitatively unchanged if we instead consider all trees (including singletons).

It is important to note that there are two distinct ways in which the LinkedIn signup forest is more viral than previously studied adoption structures. First, the unnormalized versions of the measures we considered above—the proportions of nodes that adopt far from root nodes, and that reside in large and deep trees—are much higher on LinkedIn. Second, even when we restrict attention to nodes explicitly involved in diffusion (either non-root nodes or nodes in non-singleton trees, depending on the analysis), the proportions are *still* significantly higher, often by an order of magnitude or more. This second difference does not necessarily follow from the first. It would have been possible for LinkedIn to be "more viral" in the sense of a larger proportion of nodes being involved in member-to-member transmission, with the transmission itself being as it was in the previous datasets. Yet this is not the case, as the conditional distributions in Figs. 2 and 3 indicate.

2.3 Structural Virality of Signup Cascades

In addition to quantifying the virality of the signup forest as a whole, we also quantify the shape of signup cascades by measuring their *structural virality*. The goal of structural virality, introduced in [13], is to numerically disambiguate between shallow, broadcast-like diffusion and the deep branching structures called to mind by the biological "viral" metaphor. The structural virality

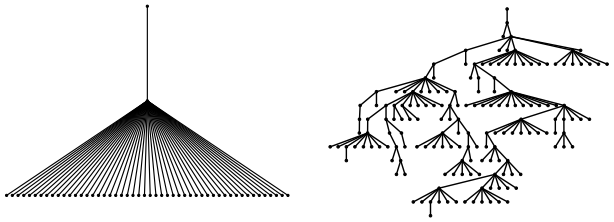


Figure 4: Two LinkedIn signup cascades, one with (left) low structural virality (Wiener index $= 1.99$), and one with (right) high structural virality (Wiener index $= 9.5$).

measure, called the Wiener index, is equal to the average path distance between two nodes in the tree. Pure broadcast diffusion (*i.e.*, a star-like cascade) results in very low scores on this measure (just under 2), while deep narrowly branching structures will have very high scores as the paths between nodes get very long (see Fig. 4 for examples of what real cascades with low and high Wiener indices look like). Low structural virality corresponds to broadcast-dominated diffusion, whereas high structural virality corresponds to multi-step transmission. In this section, we restrict our analysis to cascades with over 100 nodes, as was done in [13].

Interestingly, the distribution of structural virality over LinkedIn cascades with more than 100 nodes is qualitatively similar to the distributions reported for Twitter cascades with more than 100 nodes in [13]. Thus, although our analyses above show that the LinkedIn signup forest as a whole is significantly more viral than previously analyzed datasets, large LinkedIn signup cascades are roughly as viral as large Twitter information-sharing cascades. It's possible that LinkedIn's overall virality stems from the preponderance of large cascades, rather than the structural virality of the individual large cascades themselves, but more direct comparisons are needed to validate this hypothesis.

However, there remains an important sense in which the structural virality of LinkedIn trees follows different properties from the structural virality of the Twitter cascades in earlier work. 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 [13]. This implies that an information cascade's size does not reveal much about the structural mechanism by which it spread. In contrast, for LinkedIn signup cascades the correlation is a strikingly high 0.72, implying that the largest LinkedIn signup cascades truly spread through a viral process that is both deep and relatively narrow (Fig. 5 visualizes the structural virality of LinkedIn signup cascades as a function of their size). This high correlation appears to be related to the relative lack of broadcasts in LinkedIn: 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. As far fewer broadcast channels are available to drive adoption, sigup cascades on LinkedIn must therefore, by this definition, be more structurally viral if they are to spread to large populations. This departure from previously-studied cascades is a good example of how these differences do not follow purely from size.

Taken together, these analyses indicate that multi-step diffusion has played a much larger role in the adoption of LinkedIn than it did for the variety of domains considered in previous work. This result complements those found in previous work on information sharing cascades [13, 14, 22, 26]. Whereas they found a surprising general lack of viral propagation across a wide variety of domains,

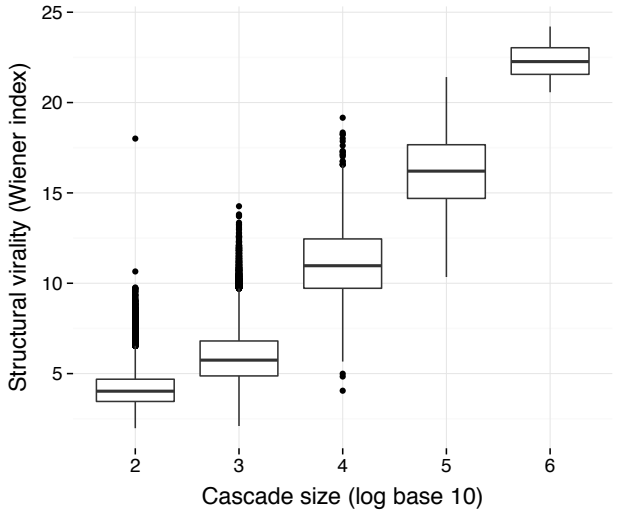


Figure 5: Structural virality as a function of cascade size (log base 10). The correlation is remarkably high, in contrast with previous findings on information-sharing cascades.

here we show that in the important case of a major global website, large-scale viral propagation does occur. We emphasize that this outcome was not preordained by size; merely because an adoption event was huge does not necessarily imply that it achieved its popularity through viral growth.

3. LOCAL AND GLOBAL HOMOPHILY

We have established that person-to-person transmission is an important mechanism through which LinkedIn has spread around the world. But what is the interplay between the diffusion structures we observe and the attributes of people involved in the diffusion process? Previous large-scale diffusion studies have largely treated users alike and concentrated on determining how—and whether—information, products, and services spread, just as we have in the previous section. As a consequence, our understanding of who is transmitting to whom (in terms of underlying user attributes) in large-scale diffusion events is very limited. Here, we connect our global signup cascades with the rich individual-level attribute data available on LinkedIn to investigate this question.

Since virtually all interpersonal networks display *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. What is less clear is how this homophily manifests itself in the composition of user attributes in the cascades, an effect that in principle can be substantial. By investigating this empirical composition, we seek to shed light on a fundamental question: are cascades more homogeneous—more *coherent*—than one would expect simply from the local level of homophily between inviters and invitees? Or is the level of homophily between inviters and invitees, when propagated over entire cascade trees, sufficient to account for the global level of homogeneity that we see in the trees as a whole?

LinkedIn is an ideal domain to study this question for two reasons: first, we have observed a high prevalence of multi-step diffusion; and second, there is a wealth of individual-level attribute data available, since most members fill in impressively detailed profiles. We have high coverage for a diversity of individual traits, such as country of residence, geographic sub-region, professional industry of employment, age, job type, job seniority level, and others.

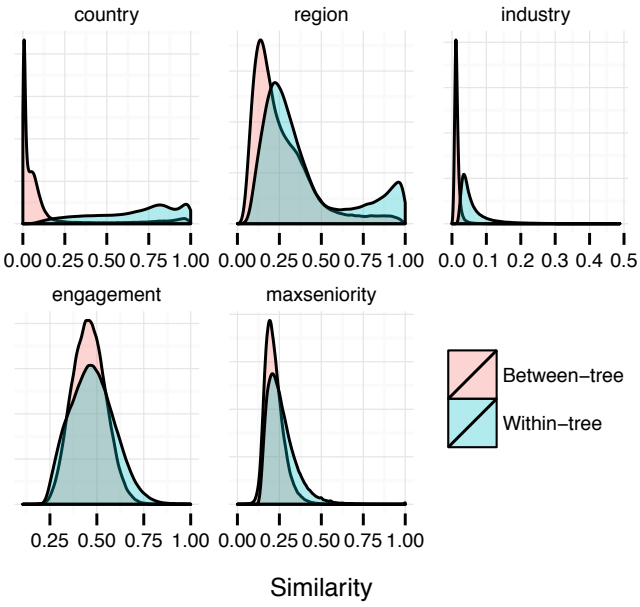


Figure 6: Within-tree and between-tree similarity on country, region, industry, engagement, and maximum job seniority.

3.1 Homophily in LinkedIn Signups

First we conduct an observational analysis of homophily in signups along various member attribute dimensions.

Edge homophily. Since homophily is such a pervasive phenomenon in interpersonal networks, we expect to see members inviting (and having their invitations accepted by) people like themselves. We can 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 $N(A_1 \rightarrow A_2)\ /\ N(A_1)$, where $N(A_1)$ is the number of signup edges originating from members with attribute value A_1, and $N(A_1 \rightarrow A_2)$ is the number of signup edges where the source and destination have attribute values A_1 and A_2, respectively.

Examining these probabilities reveals that there is indeed a significant amount of edge homophily present in our data. The "self-loop" probabilities $P(A_1|A_1)$, where members accept invitations from others like themselves to join, are much higher than the transition probabilities $P(A_2|A_1)$ between different values $A_1 \neq A_2$. For example, for the country attribute, the conditional probabilities $P(\text{Brazil}, \text{Brazil})$ and $P(\text{India}, \text{India})$ are both greater than 0.80.

Comparing these probabilities to a randomized baseline where the attribute of each node is simply drawn from the overall distribution of the LinkedIn population confirms that the real self-loop probabilities are much higher than they would be if there were no edge homophily. Thus, members invite others like themselves, beyond what we would expect from the underlying group populations.

Cascade homophily. Now we investigate the extent to which the signup *cascades* display homophily along various dimensions.

Our main object of analysis in this section is the distribution over various attributes among members within the same cascade tree. To ensure the distributions are not skewed by small-sample effects, we restrict our attention to cascades comprising at least 100 members. There are over 100,000 such cascades in our dataset.

Given a distribution over member attributes, we would like a way to quantify how similar or diverse it is. We also wish to compare two distributions, and measure how similar they are. Ideally, these two quantities should be directly comparable. We fulfill these desiderata by adopting the *population diversity* measure used in sociology [28]. We define:

- 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.

These metrics have the advantages of being easily interpretable and directly comparable, and are not affected by the size of the populations being considered.

For every "large" tree T (over 100 members) and attribute A, we compute the within-tree similarity $W_A(T)$ of the members in the tree. Then we can examine the distribution over $W_A(T)$ for every attribute A. However, as was the case with edge homophily, a large amount of cascade *similarity* is insufficient to conclude that cascade *homophily* is present. Thus, we also take a random sample of pairs of trees and calculate the between-tree similarity $B_A(T_1, T_2)$ of the two member attribute distributions. The distribution over these between-tree similarities then provides a baseline to compare against. If there were no cascade homophily on A at all, then the within-tree and between-tree similarity distributions would be exactly the same. The extent to which they differ, then, is a direct measure of cascade homophily in our data.

In Fig. 6, we show the distributions W_A and B_A over all large trees for the following attributes: country, region, industry, engagement, and maximum seniority (top job level over one's career, as reflected by the job title). There are several important points to observe. First, there is a striking amount of homophily along some dimensions. The signup cascades are extremely homophilous on the geographical attributes, especially on country: many trees have within-tree similarity values close to 1, whereas the between-tree overlap is almost always below 0.25. Industry also displays significant homophily, in that W_{industry} and B_{industry} are almost non-overlapping. Second, the extent to which homophily is present varies widely across the attributes, since there is little to no homophily on engagement or maximum seniority. Finally, the geographic attributes display an intriguing pattern: their within-tree similarity distributions are bimodal. This suggests that there are two distinct ways in which signups cascade through countries and regions, with little interpolation between the two. We will return to this fact later in this Section.

Homophily by root country. Even in the absence of homophily, members with popular attributes are more likely to be associated with each other than members with rare attributes simply because there are more of them. Thus cascades that start in the United States are more likely to display high similarity than those that start in Belgium. Here we explore how similarity and homophily change with the specific value of the root attribute, focusing on country since it displays the highest within-tree similarity.

In Fig. 7, we plot the distribution of within-tree similarity W_A for large trees rooted in Brazil, Canada, France, India, and the United States. All five countries show a high degree of similarity, the exact magnitude of which correlates with the size of country's membership on LinkedIn. The fact that a single attribute value results in such high and homogeneous similarity distributions demonstrates the remarkably strong homophily we observe on country.

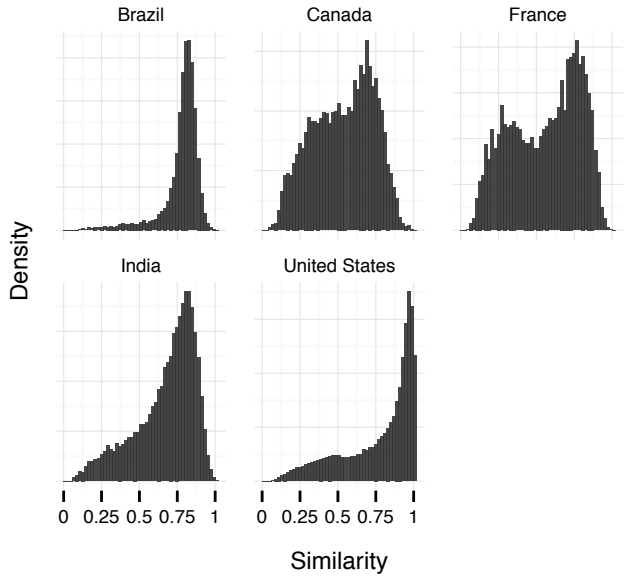


Figure 7: Within-tree similarity for trees rooted in Brazil, Canada, France, India, and the US.

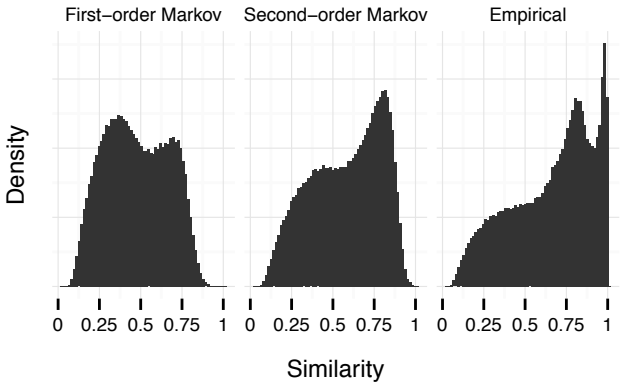


Figure 8: Within-tree similarity on real tree topologies with countries drawn from: (left) first-order Markov transitions M_1, and (middle) second-order Markov transitions M_2; (right) empirical within-tree similarity.

Also, the similarity distribution is unimodal for almost every country in our dataset. A few countries, such as France in Fig. 7, have strong bimodality on their own, but most do not; the overall bimodality we observed in Fig. 6 is related to the diversity in country size, with the resulting cascade similarity depending on where the cascade is rooted.

3.2 Levels of Homophily

We have established that there are strong edge homophily effects present in LinkedIn signups, and there are also strong cascade-level homophily effects present in the signup cascades. Already, these empirical facts have important ramifications for how the site population will evolve: given that many new members are invited to join by existing ones, and that cascades display strong homophily effects, it follows that the warm signups of tomorrow will look like the inviters of today.

However, it is unclear whether the homophily effects present in the signup cascades are different from the homophily present at a local level. Do the distributions over country of residence in cascades simply follow from the basic level of edge homophily present?

Modeling edge homophily. To explore this, we simulate a warm signup process with the same cascade topologies we observe in the data, but where member countries are drawn according to a first-order Markov chain derived from empirical data. This first-order Markov chain M_1 is defined with the conditional probabilities $P(A_2|A_1)$ computed in the previous section as transition probabilities. In our synthetic model, the member country at a given node is determined by applying one step of the empirical Markov chain transition to the country of the parent. Hence the first-order Markov chain models edge homophily, and the question is whether such a model of local edge homophily is able to reproduce the observed global homophily patterns of cascades.

We proceed as follows. For each cascade, the country of the root node is kept the same as it is in the data. Then the countries of each of the root's children are drawn independently from the Markov chain. The same is done for their children, and so on down the cascade. This induces a distribution over countries in the tree, from which we compute within-tree similarity for each cascade as before. We can then compare the distribution over within-tree similarity induced by this simple first-order Markov process with the actual distribution observed in empirical data.

From edge to cascade homophily: First-order effects. The outcome of the above process for the country attribute is shown in Fig. 8(left), in which two important patterns are immediately apparent. First, the distribution of similarity across trees is bimodal, just as it is in the empirical data (shown in Fig. 8(right)). This implies that edge homophily is sufficient to explain the bimodality in within-cascade similarity. In fact, edge homophily on a star topology (instead of the real cascade topologies) recovers the bimodality as well—thus it results from the combination of edge homophily on country and insufficient tree-depth to allow mixing to the overall country distribution.

Second, the absolute level of within-tree similarity in the Markov simulation, while still high, is significantly lower than what we observe in empirical data. A direct consequence of this is that the type of member who joins a specific invitation cascade is, on average, not entirely determined by the type of member who invites him (were that the case, the similarity patterns produced by a first-order Markov simulation would reproduce those we find empirically). For example, if a particular cascade has been spreading among members based in India, and a new member from Kuwait joins the cascade, it is more likely that this member's invitees will be from India than we would expect on average from someone from Kuwait.

Therefore, despite the strong presence of country homophily at the local (edge) level, it is insufficient to explain the country homophily we observe at the cascade level. Member attributes in cascades, then, are determined by some process above and beyond local, homophilous interactions alone, which we already found to be powerful. A new member's attributes are not governed only by her parent, but by the rest of the cascade she is a part of as well. This result illuminates a crucial point about the user composition of cascades, and answers the central question posed at the beginning of this Section: *cascades are not simply arbitrary subsets of members following global demographic correlations*—they are more coherent than this simple characterization would suggest.

From edge to cascade homophily: Second-order effects. To investigate the higher-order effects between new members and the cascades they join, we repeat the simulation above using a second-order Markov chain M_2 to generate node countries. M_2 is defined by a process analogous to the first-order case: the conditional probability that a new member with attribute value A_3 joins, given that her inviter has value A_2 and her inviter's inviter has value A_1, is $P(A_3|A_1, A_2) = N(A_1 \to A_2 \to A_3)/N(A_1 \to A_2)$, where $N(\cdot)$ again refers to the number of signup paths connecting nodes with particular attributes. If $N(A_1 \to A_2)$ is too small, then we ignore the grandparent and use the first-order probability $P(A_3|A_2)$. Using a second-order Markov chain allows us to capture effects such as in the India and Kuwait example described above.

The resulting distribution of within-tree similarity, shown in Fig. 8(middle), is shifted remarkably far to the right compared to the outcome of the first-order simulation—in this case, the "second-order effects" are actually quite large. The magnitude of the discrepancy between the first-order and second-order models shows how much more homophily structure there is at the cascade level. Furthermore, note that there is still the mode of cascades with near-perfect country similarity in empirical data that remains unaccounted for by the second-order Markov simulation, which further reinforces the coherency of signup cascades beyond local edge homophily effects.

3.3 Guessing the Root of a Cascade

The fact that the observed cascade homophily effects are not explainable via local edge homophily effects alone suggests that cascades tend to retain some "memory" of their starting point. Here we ask: how quickly does a cascade "lose" the attribute of the root node (its country, say) as the cascade grows and relaxes to the background distribution?

The process by which this happens is the subject of a well-known probabilistic model originating in the study of evolutionary trees—we imagine an attribute at the root of a tree and then this attribute is passed on to the children with some probability of mutation [12, 31, 32, 35]. In the genetic application, this attribute would be an allele of a gene, while in our case it would be some homophilous property of users, such as their country. The question is then whether, deep enough into the tree, the node attributes have mixed to some background distribution for the full population, or whether arbitrarily deep in the tree we can still infer something about the value at the root. Notice that if the tree were simply a path (*i.e.*, if each node had exactly one child), then we would expect the process to mix to the background distribution. However, on a tree with non-trivial branching factor, there are competing forces: the process tends to mix on each path, but there are many overlapping paths on which to preserve information about the value at the root.

To address this issue, we consider the following concrete "root-guessing" question for the trees in our 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? Asking how often this plurality guess correctly predicts the root's value allows us to further elucidate the extent to which global homophily is present in the trees, by seeing how often the root's characteristics can be detected deep into the tree. We will also compare the results of the root-guessing experiment on the real LinkedIn data to the results when attribute values are generated by the natural first-order or second-order Markov chains defined as above.

The results of this computation using the country attribute are shown in Fig. 9. There are a number of interesting conclusions from this experiment. First, 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, which is far beyond the maximum depth that most cascades reach. Second, the first-order Markov simulation relaxes to the global prior much faster than the empirical data does. Finally, the second-order Markov chain fares significantly better, again showing the strong higher-order homophily interactions present in signup cascades.

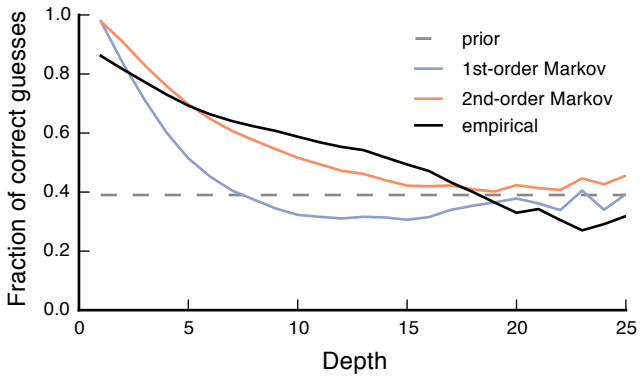


Figure 9: Fraction of time plurality attribute at depth d matches root attribute in root-guessing experiment. Empirical data retains "memory" of the root longer than baselines.

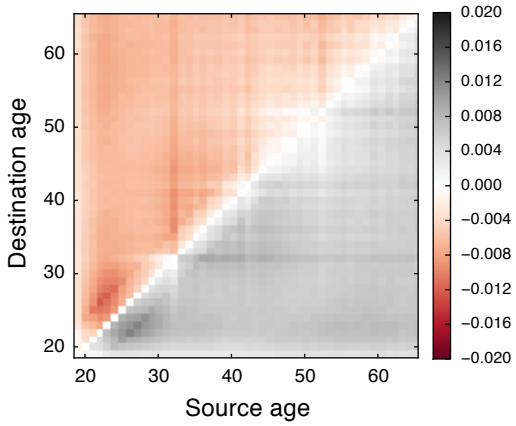
It is interesting to think about the role of the second-order Markov chain in light of the analogy to the genetic applications of the model. In genetic contexts, the process is a first-order Markov chain by the definition of genetic inheritance: conditional on knowing the true complete genotypes of an individuals' parents, there is in essence no additional information contained in the genotypes of the grandparents. But in our social setting, the country of a node's parent in the tree no longer serves as a sufficiently complete description—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. In effect, a small amount of profile information may be serving more as a kind of "social phenotype" rather than a "social genotype," displaying only observable characteristics rather than deeper internal ones, and the use of the higher-order Markov chains may help fill in some of the missing information that results.

3.4 Status Gradients

Throughout this section, we've discussed how signup cascades show strong patterns of homophily along certain attributes like country and industry, whereas along other attributes signups aren't homophilous. However, some of these other attributes, such as age and job seniority, show other structure.

As in offline realms of professional life, *status* is an important part of one's identity on LinkedIn. Signup edges are inherently directed: one member issues an invitation and the other member joins a community through that person. Thus 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.

We check the extent to which this effect occurs empirically. In Fig. 10(a), 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 $P(y|x) - P(x|y)$, where $P(u|v)$ means the probability that a member of type u accepts an invitation, given that it originated from a member of type v). There is a clear effect on age: the grey below the $x = y$



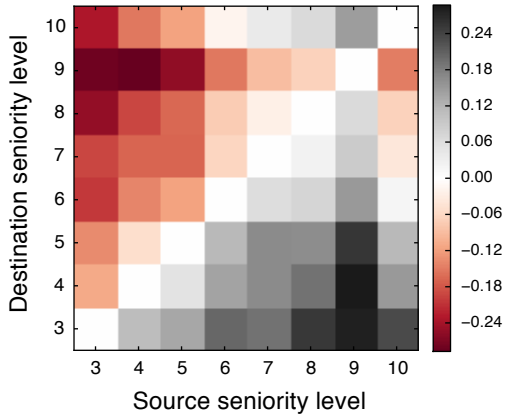


Figure 10: Status gradients on age and maximum seniority.

diagonal indicates that younger members are more likely to accept invitations from older members than vice versa, indicating there is a status effect on age governing who invites whom and who accepts whom's invitations. In Fig. 10(b), we 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, we define a member's seniority to be the highest level they've ever worked at).

Thus there are two fundamental ways signups flow through networks: along certain attributes, members tend to act homophilously and invite others like themselves, and along others, signups tend to progress down status gradients, flowing from higher-status users to lower-status users.

4. CASCADE GROWTH

Having considered the signup cascades as static objects, we now trace their development over time and investigate various aspects of their temporal evolution.

4.1 Timescales of transmission

A key characteristic of any diffusion process is how much time elapses between adjacent adoptions. In the biology-inspired terminology often employed in diffusion models, when does an "infected" node transmit the contagion (*i.e.*, the signup) to another?

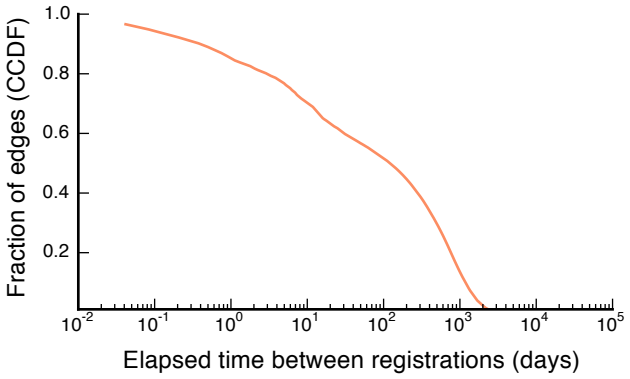


Figure 11: Complementary cumulative distribution function (CCDF) of elapsed times between inviter and invitee signup times. Adoptions are usually very separated in time.

To answer this question, we 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. Then we examine the distribution of elapsed times between when A joined and when B accepted the invitation from A. Fig. 12 shows this distribution for members who joined in 2006 (all time frames are qualitatively similar). It is immediately apparent that adjacent adoptions of LinkedIn are widely separated in time: around 40% of members who joined did so at least a year later than their inviters did. This is in contrast with other diffusion settings, such as information-sharing on Facebook or Twitter, where the majority of transmissions have been observed to occur within a few days of adoption [11, 39].

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, or invitees could be accepting invitations long after they receive them. We check this directly: Fig. 12 shows the fraction of invitations sent as a function of time for users who joined in March 2012 (other times were qualitatively similar). We 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 (this latter fact is illustrated in Fig. 13).

Based on these results, we conclude that members of LinkedIn remain "infectious"—able to participate in member-to-member diffusion—over very long periods of time; it is not the case that the majority of transmissions occur in some narrow time frame relative to adoption. Extremely long infectious periods can clearly contribute to the success of a cascade, since members can invite others to the network during a wide range of times.

4.2 Cascade growth trajectories

This fact has a simple but important consequence: if individual edges often take months or years to form, then the cascades they make up must persist for long periods of time as well. In the faster-paced context of online social media, the adoption of popular pieces of online content is often concentrated within a narrow time frame [23, 39], and sharing cascades consequently achieve much of their growth during this small interval. But this is unlikely in our setting given how much longer signup cascades persist. Thus, we ask: How do LinkedIn cascades grow in size over time?

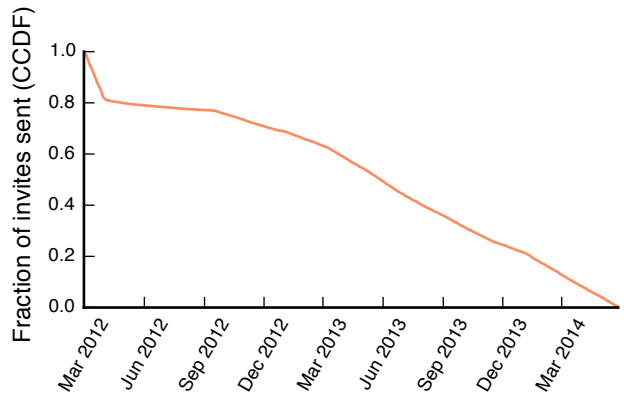


Figure 12: CCDF of the number of invites sent as a function of time for members who joined in March 2012. Most invitations are sent months after a member joins, meaning members remain "infectious" over very long periods of time.

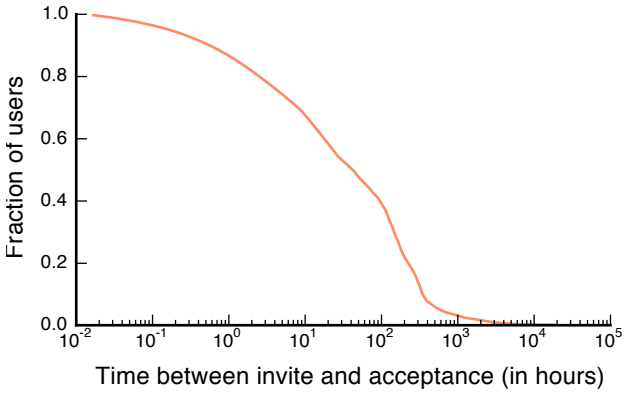


Figure 13: CCDF of the elapsed time between when an invitation was sent and when it was accepted. Invitations are accepted very quickly after they are sent.

To answer this question, we plot the growth trajectory of the 1,000 biggest cascades on LinkedIn in Fig. 14(a). For each cascade, we normalize both time and size to be between 0 and 1, and show the fraction of the cascade's eventual size at various points in time between the root's registration and the present day. A surprisingly robust linear growth pattern is apparent. Although in principle cascades could have reached their eventual size in very different ways, this doesn't happen in practice. There is very little variation in how big cascades became big on LinkedIn; virtually all of them gradually accumulated members at a constant rate over time. Thus, LinkedIn's rapid growth is not accounted for by individual cascades alone—it is the number of distinct cascades growing in parallel, each of which is growing relatively linearly, that increases so dramatically. In Fig. 14(b), the same type of plot is shown for 1,000 medium-sized cascades. Much more variation is present, but it is still around the same basic linear trend.

We conclude that, in contrast with the intuitive (and largely accurate) picture of viral videos, pictures, and news stories quickly spreading through online media and interpersonal networks before mostly burning out, the picture of diffusion that emerges from our study of LinkedIn is one of persistent, parallel accumulation of subpopulations at a much more deliberate pace.

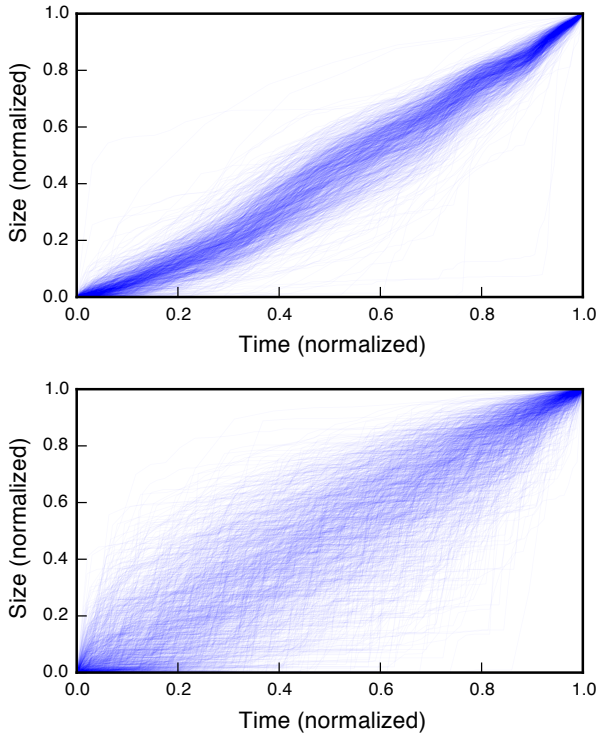


Figure 14: Patterns of cascade growth over time for (top) large cascades (over 4,000 members) and (bottom) medium cascades (500 members). Cascades grow relatively linearly over time, bigger cascades more consistently so than smaller ones.

Comparison with random baseline. The global picture of LinkedIn's growth involves many trees accumulating nodes in parallel at a notably linear rate. It's natural to ask whether a simple generative process of tree growth can reproduce the basic properties we've observed: the distribution over tree sizes, as well as the linear growth rate pattern.

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. The choice of how nodes are assigned to be warm or cold is thus the only parameter in the process, and for this purpose we use the real exact ordering of warm and cold signups over the history of LinkedIn. The resulting distribution of cascade sizes is remarkably robust: 30 runs of this process are plotted in color in Fig. 15, and the empirical distribution of cascade sizes is shown in black (we only show non-trivial cascades, as in Section 2).

Relative to the real distribution, the slopes of the distribution for draws from the randomized baseline are roughly the same, but they are consistently shifted left: the individual trees are larger in the real distribution from LinkedIn.

To determine if this process gives linear growth trajectories for individual trees, we add timestamps to this model. We consider the same simple process as above, in which each node arrives with its true warm/cold status and now also its true arrival time, and randomly draws its parent uniformly at random from all existing nodes. We find that the rates of tree growth follow a linear trend very close to the real empirical distribution shown in Fig. 14, and with a variance across trees that is even smaller than in the real data.

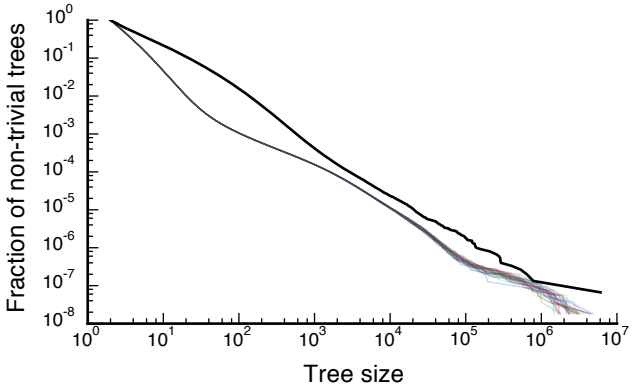


Figure 15: Distribution (CCDF) of cascade sizes on random baseline with real ordering of warm and cold signups (color) and empirical distribution of cascade sizes (black).

5. RELATED WORK

There are three lines of research related to our work here: the dynamics of cascades; the growth and evolution of online social networks; and analyses of homophily, the tendency of people to connect to others who are similar to them.

Cascades. Our work builds on the rich literature that studies diffusion and adoption of new ideas, products, and behaviors [33]. Early cascade research in the social science literature was based on focused empirical studies [9] as well as mathematical models [6, 16, 38]; only relatively recently have scientists have able to observe and measure large-scale diffusion events. Online cascades served as sources of detailed data about such events; studies drawing on this type of data have considered settings such as blogging [1, 17, 26], e-mail [15, 27], product recommendations [22], and the sharing of information in social sites such as Facebook, Flickr, and Twitter [7, 8, 11, 13, 14, 19, 21]. In work that aligns more closely with our focus on signup cascades, researchers have also studied cascades of group formation [4] and adoption of online services [3, 5, 34, 36]. Much of this previous work has the property that even the successful cascade events in the data were much smaller in scale [8, 22] than the signup cascades we consider here.

We compare the structure of LinkedIn's signup cascades with several datasets from [14]. In that work, Goel et al. characterized the structure of online diffusion networks in several domains, and observed that most cascades on the Web are shallow and small. In contrast, we find that LinkedIn signup cascades tend to be deeper and larger, and grow steadily over long periods of time. Our case study of the LinkedIn signups forest, an extreme example of online adoption, complements the general picture of online diffusion structures drawn in [14] by filling in what one of the largest online diffusion structures looks like. In addition, we apply the structural virality measure introduced in [13] and find that the correlation between size and structural virality is very high for LinkedIn signups, in contrast to the very low correlations found for information-sharing cascades in [13].

Growth and evolution of social networks. A second line of related work is on mechanisms for the evolution of online social networks, using data covering the growth of social networking services including Facebook [37], Flickr, [20, 30], LinkedIn [24], and others [25]. This line of work investigates the evolution of network structure assuming an underlying arrival process for new nodes. Our work, on the other hand, focuses on the mechanisms that underpins this arrival process, through the new members who join the network and the ways in which cascades of invitations spread.

Homophily in social networks. Homophily—the tendency of people to associate with others like themselves—is one of the most important forces shaping the structure of social networks [29]. Recent work using online data has considered the challenges in differentiating homophily from influence [2, 10, 34], and has established links between the evolution of social network structure and the emergence of homophily [18]. We focus here on a distinct issue that is particularly well-suited to analysis via our signup cascade data—the ways in which local patterns of homophily between pairs of individuals translate into more global forms of homophily at the level of an entire cascade, and how cascade-level homophily can transcend homophily observed at local levels.

As we discuss in Section **??**, part of our analysis of this local-global link involves developing an intriguing connection with a body of mathematical work that has developed separately from the homophily literature—namely, research on phylogenetic tree reconstruction [12, 31]. That problem is formulated in terms of a process in which information (*e.g.*, a binary attribute) is recursively transmitted from a root node down a tree. At every step the attribute can mutate with some given probability, and the goal is to reconstruct the attribute of the root given the values of the attributes of all the children at depth d [32, 35]. The connection we develop begins with the observation that a homophilous attribute in a cascade can be thought of as analogous with a genetic trait that propagates down an evolutionary tree, changing to a new value with some probability. The extent to which genetic information can be reconstructed about ancestors then turns into the question of inferring properties of a cascade's initial starter from the properties of its eventual adopters. We show how this inference depends intimately on the structure of the cascade tree and the homophily of the attribute, and we argue that the type of Markov modeling assumption needed in the social context appears to differ from what is relevant in the genetic context.

6. CONCLUSION

By analyzing the global spread of LinkedIn, we 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. We 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. We 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.

Several points from our earlier discussions are worth drawing out in greater detail. First, while the cascading adoption of LinkedIn has reached a very large user population, it is important to emphasize that the structural properties of its spread are not due to scale alone. Indeed, recent work demonstrated that the virality of the cascade trees can be roughly independent of their size [13]. The massive signup cascades we study here show that very different kinds of diffusion are possible—in which virality increases with tree size, coupled with persistent linear tree growth over time.

Second, the interaction between local and global similarity in the cascade trees points to deeper issues about the nature of homophily. While we tend to think of homophily patterns as arising from the accumulation of local similarity along the links of a social network, our Markov-chain analysis shows that the cascade trees exhibit a higher level of global similarity than would follow from

these purely local effects—members of signups cascades have a certain coherence to them that is not explainable via simple models of pairwise interaction. One possibility, suggested by the analogies to genetic models and the corresponding limitations we identify in these analogies, is that one needs a richer type of "social genome" to characterize each individual in the cascade—in essence, a profile detailed enough that knowing the profile of an individual's parent provides sufficient information to estimate properties of the individual. Identifying the bases for such profiles, and for the mechanisms by which these characteristics propagate through social networks, could provide a new way of reasoning about the processes by which large groups on the Web come together and adopt new products and innovations.

Acknowledgments. We thank Sam Shah and Myunghwan Kim of LinkedIn for their help. This research has been supported in part by a Google PhD Fellowship, a Simons Investigator Award, a Google Research Grant, a Facebook Faculty Research Grant, an ARO MURI grant, NSF grants IIS-1016909, IIS-1149837, IIS-1159679, CNS-1010921, IIS-0910664, Boeing, Facebook, Volkswagen, and Yahoo.

7. REFERENCES

[1] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. In *Workshop on the Weblogging Ecosystem*, 2004.

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *Proc. KDD*, 2008.

[3] S. Aral and D. Walker. Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science*, 2011.

[4] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: Membership, growth, and evolution. In *Proc. KDD*, 2006.

[5] E. Bakshy, B. Karrer, and L. A. Adamic. Social influence and the diffusion of user-created content. In *Proc EC*, 2009.

[6] D. Centola and M. Macy. Complex contagions and the weakness of long ties. *American Journal of Sociology*, 2007.

[7] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi. Characterizing social cascades in flickr. In *Proc. WOSN*, 2008.

[8] J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec. Can cascades be predicted? In *Proc. WWW*, 2014.

[9] J. Coleman, E. Katz, and H. Menzel. The diffusion of innovation among physicians. *Sociometry*, 1957.

[10] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *Proc. KDD*, 2008.

[11] P. A. Dow, L. A. Adamic, and A. Friggeri. The anatomy of large facebook cascades. In *Proc. ICWSM*, 2013.

[12] W. Evans, C. Kenyon, Y. Peres, and L. J. Schulman. Broadcasting on trees and the ising model. *Annals of Applied Probability*, 2000.

[13] S. Goel, A. Anderson, J. Hofman, and D. Watts. The structural virality of online diffusion. *Under review*.

[14] S. Goel, D. J. Watts, and D. G. Goldstein. The structure of online diffusion networks. In *Proc. EC*, 2012.

[15] B. Golub and M. O. Jackson. Using selection bias to explain the observed structure of internet diffusions. *Proc. Natl. Acad. Sci.*, 2010.

[16] M. S. Granovetter. Threshold models of collective behavior. *American Journal of Sociology*, 1978.

[17] D. Gruhl, R. V. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proc. WWW*, 2004.

[18] G. Kossinets and D. J. Watts. Origins of homophily in an evolving social network. *American Journal of Sociology*, 2009.

[19] R. Kumar, M. Mahdian, and M. McGlohon. Dynamics of conversations. In *Proc. KDD*, 2010.

[20] R. Kumar, J. Novak, and A. Tomkins. Structure and evolution of online social networks. In *Proc. KDD*, 2006.

[21] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proc. WWW*, 2010.

[22] J. Leskovec, L. Adamic, and B. Huberman. The dynamics of viral marketing. *ACM Transactions on the Web*, 2007.

[23] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *Proc. KDD*, 2009.

[24] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *Proc. KDD*, 2008.

[25] J. Leskovec, J. M. Kleinberg, and C. Faloutsos. Graph evolution: Densification and shrinking diameters. *ACM Transactions on the KDD*, 2007.

[26] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *Proc SDM*, 2007.

[27] D. Liben-Nowell and J. Kleinberg. Tracing information flow on a global scale using Internet chain-letter data. *Proc. Natl. Acad. Sci.*, 2008.

[28] S. Lieberson. Measuring population diversity. *American Sociological Review*, 1969.

[29] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 2001.

[30] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the Flickr Social Network. In *Proc. WOSN*, 2008.

[31] E. Mossel. Phase transitions in phylogeny. *Transactions of the American Mathematical Society*, 2004.

[32] E. Mossel, S. Roch, and A. Sly. On the inference of large phylogenies with long branches: How long is too long? *Bulletin of mathematical biology*, 2011.

[33] E. M. Rogers. *Diffusion of Innovations*. Free Press, New York, fourth edition, 1995.

[34] A. S. Sinan Aral, Lev Muchnika. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl. Acad. Sci.*, 2009.

[35] A. Sly. Reconstruction for the potts model. In *Proc. STOC*, 2009.

[36] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg. Structural diversity in social contagion. *Proc. Natl. Acad. Sci.*, 2012.

[37] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the Evolution of User Interaction in Facebook. In *Proc. WOSN*, 2009.

[38] D. J. Watts. A simple model of global cascades on random networks. *Proc. Natl. Acad. Sci.*, 2002.

[39] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proc. WSDM*, 2011.