


Investigating and modeling the dynamics of long ties

Ding Lyu¹ , Yuan Yuan^{2,3✉}, Lin Wang¹, Xiaofan Wang^{1,4} & Alex Pentland^{3,5}

Long ties, the social ties that bridge different communities, are widely believed to play crucial roles in spreading novel information in social networks. However, some existing network theories and prediction models indicate that long ties might dissolve quickly or eventually become redundant, thus putting into question the long-term value of long ties. Our empirical analysis of real-world dynamic networks shows that contrary to such reasoning, long ties are more likely to persist than other social ties, and that many of them constantly function as social bridges without being embedded in local networks. Using a cost-benefit analysis model combined with machine learning, we show that long ties are highly beneficial, which instinctively motivates people to expend extra effort to maintain them. This partly explains why long ties are more persistent than what has been suggested by many existing theories and models. Overall, our study suggests the need for social interventions that can promote the formation of long ties, such as mixing people with diverse backgrounds.

¹Department of Automation, Shanghai Jiao Tong University, Shanghai, Shanghai, China. ²Krannert School of Management, Purdue University, West Lafayette, IN, USA. ³Connection Science, Massachusetts Institute of Technology, Cambridge, MA, USA. ⁴Department of Automation, Shanghai University, Shanghai, Shanghai, China. ⁵Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA. ✉email: yuan yuan@purdue.edu

Social network analysis provides a powerful instrument to investigate the structure of society by aggregating interpersonal relationships among individuals^{1–5}. In the social network literature, a large body of research centers on how tightly clustered social ties and groups are formed, as well as how they evolve, spread information and behaviors, and promote group solidarity^{6–12}. Meanwhile, a smaller but increasing number of studies focus on weak ties, which may function as “bridges” between different communities because of the unique roles they play in global network structures and information diffusion^{1,13–20}.

One recent development in the literature is the concept of “long ties.” These are social ties that have a large tie range, which is measured by the length of the second shortest path between two connected nodes (see Fig. 1). Long ties—social ties with a large tie range—work as important social network bridges between different communities^{21–26}. Structurally, long ties may be considered to be weak ties, as they are not positioned in a “cohesive embedded network” where individuals can easily contact or spend time with common neighbors^{14,22,27}. Yet, despite the seeming weakness (in terms of low frequency or intensity of contact) of long ties, many studies have shown that long ties are crucial for the widespread dispersion of novel information and contagious behaviors^{1,14,18,25,28–30}. Relatedly, these bridges may have other special characteristics such as exhibiting a higher level of direct reciprocity³¹.

Still, one crucial perspective lacking in the literature of long ties is the dynamics. Evidence from static social networks may not be generalizable to dynamic networks³². In particular, existing social network theories and prediction models may indirectly imply that long ties should dissolve quickly or eventually become redundant, thus putting into question the long-term value of long ties.

The critical role of long ties would be challenged if empirical evidence from dynamic networks suggests that long ties tend to dissolve or become short ties. Firstly, it is possible that long ties may dissolve rapidly. According to various theories^{14,27} and prediction models^{9,33}, social ties are likely to dissolve quickly when they lack sufficient common neighbors to reinforce their relationships or when they have few interactions (i.e., interactions with weak tie strength). Long ties likely satisfy this condition, and thus their role in bridging different communities might be limited¹⁶. Secondly, long ties may evolve to become redundant “short ties.” By triadic closure^{33,34}, a person may introduce other friends to their long ties, thereby forming common neighbors and switching the long tie to a short tie. Therefore, two people who had a long tie may become increasingly similar, for example, regarding the information they digest or the opinions they hold³⁵. Eventually, the previously long tie becomes largely redundant, as there now exist other paths where the same piece of novel information can flow between the two individuals^{27,36}.

Our study combines empirical analysis and computational modeling to provide a dynamic perspective of long ties.

First, using 2-year social network data, we find that contrary to what is implied by existing theories and models, not only are long ties more likely to persist than shorter-range ties but also that many of them continue to be long ties. To explain this finding, we propose three possible hypotheses: degree heterogeneity, survival bias, and valuable long ties^{37,38}. Investigating these hypotheses, we empirically show that the first two mechanisms might not fully explain our main results.

Next, we propose a cost-benefit analysis model to support our last hypothesis—that individuals spend extra effort to maintain relationships with long ties because they are highly beneficial, since they provide novel information or different expertise. The model combines strategic network formation models from the game theory literature^{3,7} and node embedding techniques in machine learning^{39–41} to simulate the dynamics of social networks. This interdisciplinary approach has been shown effective in trading off the model’s power to explain mechanisms versus to predict⁴². Our model describes the social tie formation process as a result of a meeting procedure and a subsequent rational decision procedure. We verify the model by utilizing real-world data. Ultimately, we find that our model partly explains the persistency of long ties, which is the main conclusion of our empirical analysis.

Results

Long ties last longer. In this work, we employ tie range to characterize the local network structure of a social tie. As the length of our data is two years, we partition the data into eight phases; our results are robust to other ways of partitioning, as well (see Supplementary Note 2). To begin our analysis, we classify all social ties by tie range in the first phase, and then, we observe the evolution of those ties in the subsequent phases.

First, we examine the dynamics of tie strength, which is measured by interaction frequency (the number of calls or texts) and interaction duration (the total duration of the calls). We define y_t as the interaction frequency or duration in phase t . We present $E[y_t | y_1 > 0]$ in the eight phases, as shown in Fig. 2. This conditional expectation indicates that we focus our analysis on ties that already exist in phase 1 (see Supplementary Note 4). Observing the magnitudes in just the first phase, we find a “U-shape” in the data that is consistent with the results of the prior work²⁵. Our result shows that interaction frequency and duration initially decrease with the tie range, but later increase with the tie range. In particular, long ties (tie range ≥ 6) appear to be as intimate as those short ties with tie range = 2 in that the average interaction frequency or duration for these two types of ties are close in the first phase.

By comparing the dynamics of short ties and long ties in Fig. 2, we find that long ties continue to be stronger. For example, in the long run, the average interaction duration and frequency of social ties with a tie range ≥ 6 appear to be even slightly larger than those

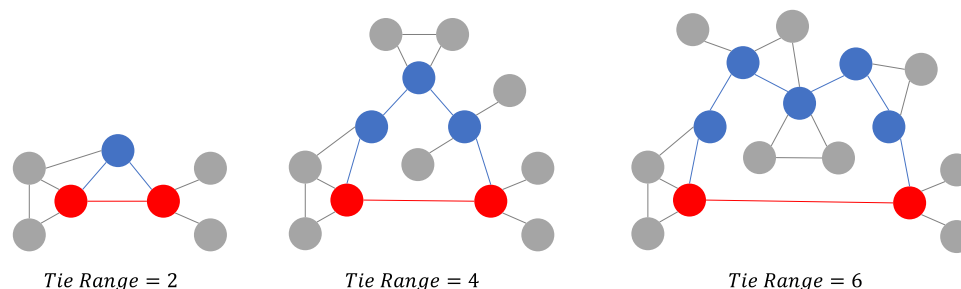


Fig. 1 Definition of tie range. Tie range characterizes the length of the second shortest path between two connected nodes. The blue nodes are the nodes on the second shortest path between the two red nodes.

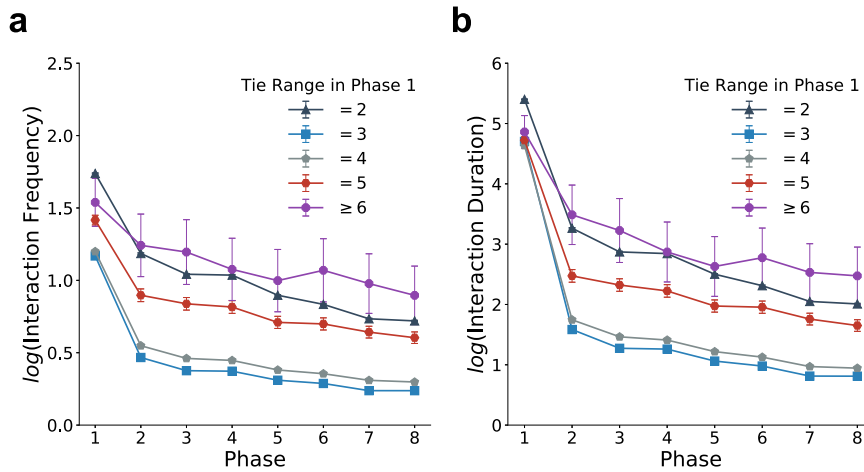


Fig. 2 Dynamics of tie strength given initial tie range. Tie strength is measured by interaction duration (the total call volume in seconds) and interaction frequency (the number of calls or texts). Each phase represents a season (three months). We take logarithms (log) for both interaction duration and frequency. All ties are classified according to their tie range in the first phase. The curves represent (a) the average (log) interaction frequency and (b) the average (log) interaction duration conditional on a tie existing in phase 1 with the given tie range. Error bars are 95% confidence intervals for the mean log interaction duration and frequency (assuming normal distribution). Note that error bars are sometimes smaller than the data point markers.

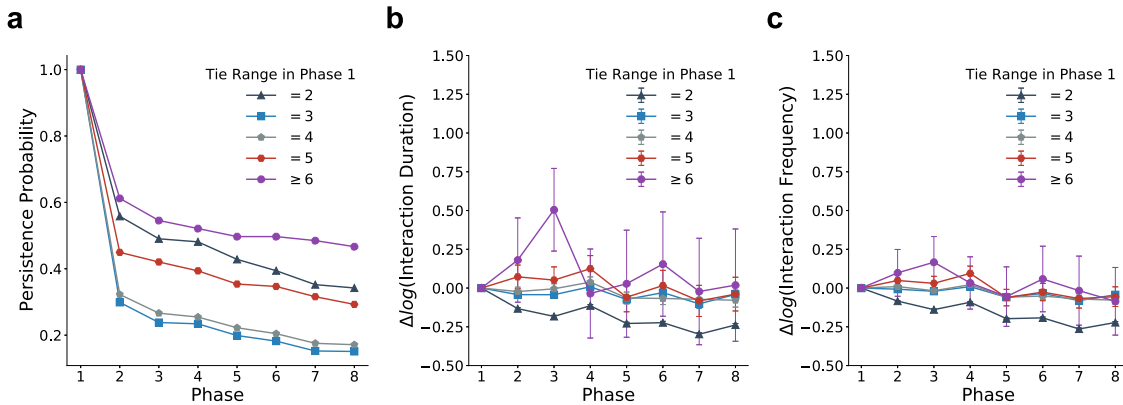


Fig. 3 Dynamics of persistence probability and interaction increments given initial tie range. Each phase represents a season (3 months). All ties are classified according to their tie range in the first phase. The curves represent (a) the probability of persisting, (b) the average ($\Delta \log$) interaction duration, and the average ($\Delta \log$) interaction frequency (c) conditional on a tie existing in phase 1 with the given tie range. Error bars are 95% confidence intervals for the means (assuming normal distribution). Note that error bars are sometimes smaller than the data point markers.

with a tie range of 2. Furthermore, social ties with a tie range of 5 also appear to be stronger than ties with a tie range of 3 or 4. In Supplementary Note 2, we discuss the robustness of our findings by adjusting the time window that determines the length of each phase.

To understand what mechanisms drive the patterns above, we decompose the dynamics of interaction frequency or duration into persistence probability and interaction increments. We let the difference in the interaction frequency or duration between phase t and 1 be $\Delta y_t = y_t - y_1$. Then, we define the persistence probability and interaction increments as follows:

$$\begin{aligned} \mathbb{E}[y_t | y_1 > 0] &= \mathbb{E}[y_1 + \Delta y_t | y_t > 0, y_1 > 0] \mathbb{P}[y_t > 0 | y_1 > 0] \\ &= \left(\underbrace{\mathbb{E}[y_1 | y_t > 0, y_1 > 0]}_{\text{persistence probability}} + \underbrace{\mathbb{E}[\Delta y_t | y_t > 0, y_1 > 0]}_{\text{interaction increments}} \right) \mathbb{P}[y_t > 0 | y_1 > 0] \end{aligned} \quad (1)$$

The dynamics of the persistence probability and interaction increments are presented in Fig. 3. As illustrated in the left panel of this figure, we find that social ties with a tie range ≥ 6 have the

largest persistence probability in all subsequent phases, followed by closely embedded ties with a tie range of 2. Meanwhile, we find that social ties with a mid-sized tie range (i.e., 3 or 4) dissolve the fastest. This pattern is consistent with the overall effect presented in Fig. 2. In Supplementary Note 5, our additional analysis show that in general, long ties have longer lifespans. Note that when defining the lifespan, we explore two choices: (1) the social tie has to have interactions for every phase within the lifespan; and (2) a social tie has interactions in the first and the last phases no matter whether they have interactions in the phases in between. The latter considers the ties being re-established after termination. The conclusion does not change with the choice of the definition of lifespan (see Supplementary Note 5). These results also show that long ties tend to be persistent longer overtime.

Regarding the interaction increments, we find that they generally increase with tie range. This indicates that conditional on a persistent social tie, the interaction frequency and duration appear to be larger when there is a long tie. By contrast, social ties with a tie range of 2 have the smallest interaction increments. From this, we conjecture that persistent short ties typically require less effort to maintain, as they can be indirectly maintained through their common friends; by contrast, we

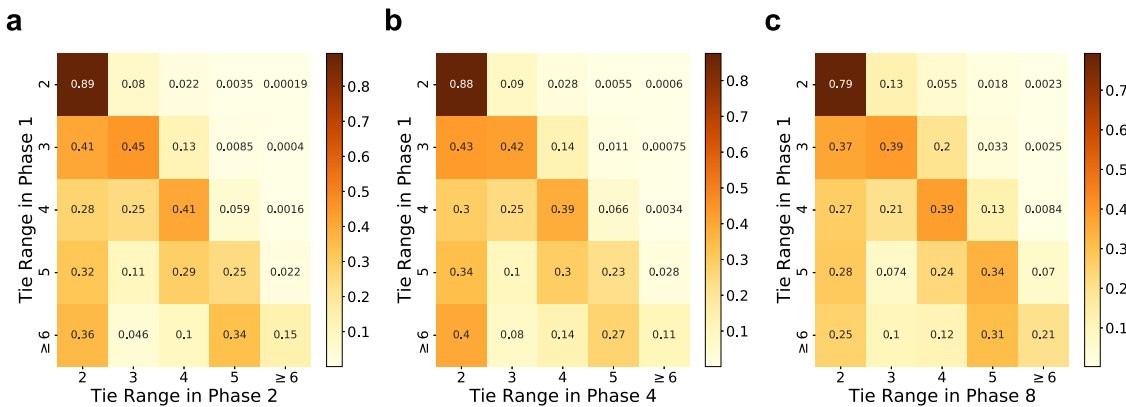


Fig. 4 Transition probability matrix of tie range from this to the next phase. The y-axis and x-axis represent tie range of social ties in phase 1 and in a subsequent phase, respectively. Social ties that dissolved in the corresponding phase are disregarded in the analysis. The numbers on the cells indicate the corresponding transition probabilities from phase 1 to (a) phase 2, (b) phase 4, and (c) phase 8.

speculate that long ties require a lot of time investment in order to be maintained.

Many long ties are persistently long. Next, we investigate the dynamics of tie range. We first examine the dynamic trends of tie range in the first two phases by analyzing the social ties that exist in both phases. We present the transition probability matrix between tie ranges in the left panel of Fig. 4. As shown in the figure, all social ties have a large likelihood of evolving into short ties. In particular, for longer ties, i.e., those with a tie range of ≥ 5 or ≥ 6 , their probability of evolving into a tie range equal to 2 is the largest: 32% or 36%, respectively. Few short ties become long ties, since such an evolution requires that all their common neighbors dissolve with either of them. In addition, long ties appear to be a stable status. For example, a social tie range ≥ 6 in phase 1 has a probability of 34% or 15% to have a tie range of 5 or ≥ 6 in phase 2, respectively.

We further analyze the tie range dynamics in phase 4 and phase 8, which are presented in the middle and right panels of Fig. 4. We find the patterns in phases 4 and 8 are largely consistent with the pattern in phase 2. In particular, for those with a tie range = 5 or ≥ 6 in phase 1, they have a probability of 26% or 38%, respectively, to persist with a tie range ≥ 5 in phase 4; they also have a probability of 41% or 52%, respectively, to persist with a tie range ≥ 5 in phase 8. These results indicate that although long ties have a high probability of becoming short ties, they can also persist as long ties. This finding suggests that it is not necessary for a social tie to become a short-range tie to be long-lasting.

Next, we proceed to jointly investigate tie range and tie strength (i.e., the frequency and the total duration of interactions). As shown in Fig. 5, in general, those ties that become short-range (e.g., tie range = 2) are those with more interactions; for social ties that have an arbitrary initial tie range but later change to a tie range of 2, the interaction frequency and duration are always the greatest. For the persistence probability, the same trend generally holds. The one exception here is for those with a tie range ≥ 6 : if they continue to be social ties with a tie range ≥ 6 , their tie strength remains strong. Note that although we are only discussing phase 1 and phase 2, our results are equally robust when we examine any phase t and its first subsequent phase, $t + 1$ (see Supplementary Fig. S9).

Explaining the results: three hypotheses. In the previous sections, we show that long ties are not only stronger but also last longer. Moreover, quite a few strong long ties continue to be long

ties. To discuss the plausible explanations for the observed patterns, We next propose and discuss three hypotheses pertaining to degree heterogeneity, survival bias, and valuable long ties below.

Degree heterogeneity. First, one plausible explanation for the observed patterns is degree heterogeneity. As shown in Supplementary Fig. S10, we find that individuals who have fewer friends are more likely to have long ties. Thus, they tend to retain relationships with a small number of friends, but with greater tie strength.

To reduce the impact of degree heterogeneity, we plot the results conditional on the degree subgroup (see Supplementary Note 6). Specifically, we separate individuals by their degree and obtain multiple degree subgroups. We then plot the main results for each degree subgroup in Supplementary Fig. S11. We find that the patterns observed in our main text are found in all degree subgroups. This finding shows that although degree heterogeneity may provide an explanation for the observed patterns, it does not fully explain our main results.

Survival bias. The second plausible explanation is survival bias—that only very valuable long ties survived—even though newly formed long ties are likely to be weaker than newly formed short ties. Therefore, surviving long ties tend to continue to persist, or perhaps even become stronger, while others dissolve rapidly. To test this hypothesis, we need to examine (1) whether newly formed long ties are weaker than newly formed short ties in the beginning and (2) whether newly formed long ties have a smaller persistence probability, such that only very strong long ties survive. We find that while (1) is supported, (2) is not supported; thus, survival bias cannot fully explain our results.

To investigate these two ideas, we divide social ties into one of two categories: existing ties, and new ties. An existing tie is one that has had any interactions in the previous phase, while a new tie has had no such interactions. After separating all ties into existing or new ones, we perform the same analysis as that found in the previous sections. We use the tie range in phase 2 as the reference, and we investigate whether there was non-zero interaction frequency or duration in order to determine if it is a new or existing tie.

We first examine whether newly formed long ties are weaker initially than newly formed short ties. In Fig. 6, we show that while existing ties present a “U-shape” in the relationship between interaction frequency (duration) and tie range in phase 2, this “U-shape” pattern does not hold for new ties. Instead, as

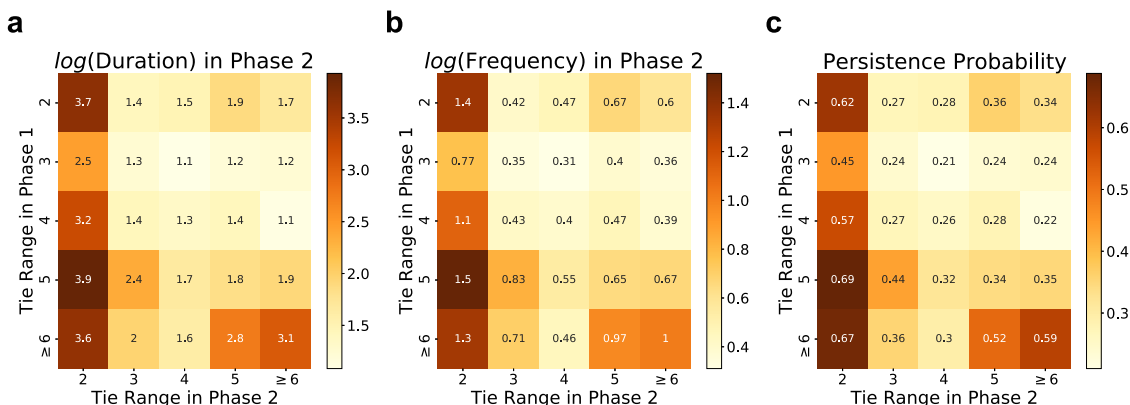


Fig. 5 Interaction duration, frequency, and persistent probability in the subsequent phase when tie range evolves. The y-axis and x-axis represent tie range of social ties in phase 1 and in phase 2, respectively. Interaction duration is measured by call volume in seconds. Interaction frequency is the number of calls or texts. Persistence probability is defined as the probability of social ties persisting from phase 1 to phase 2. The numbers on the cells indicate (a) the log means of interaction duration, (b) the log means of interaction frequency, and (c) the probability of persisting in the next phase.

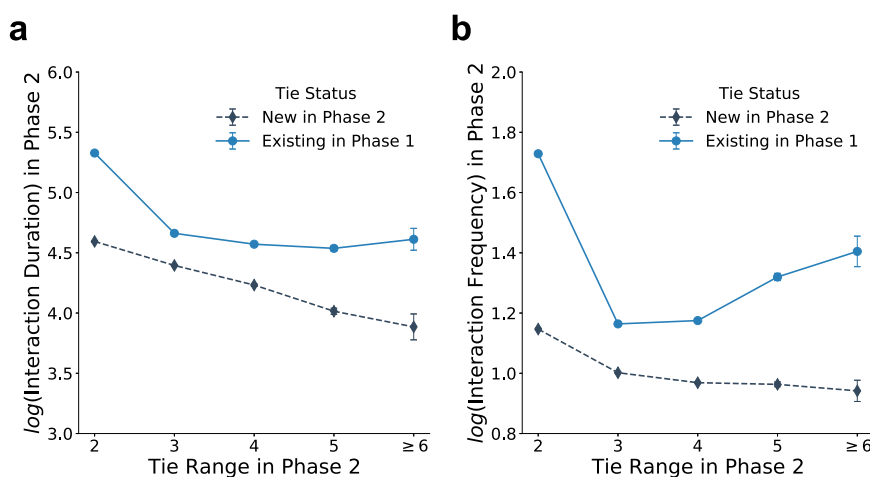


Fig. 6 Interaction duration and frequency for newly formed ties and existing ties conditional on an existing tie in phase 2. An existing tie is one that has had any interactions in the previous phase, while a new tie has had no such interactions. All ties are classified according to their tie range in phase 2. The curves represent (a) the average (log) interaction duration and (b) the average (log) interaction frequency in phase 2 of newly formed ties and existing ties with the given tie range, respectively. Error bars are 95% confidence intervals for the means (assuming normal distribution). Note that error bars are sometimes smaller than the data point markers.

indicated by Fig. 6, for new ties, the longer the new tie is, the fewer interactions the two people have in phase 2. This result supports our conjecture that newly formed long ties are likely to be weaker than newly formed short ties.

Next, we investigate whether newly formed long ties have a smaller persistence probability. However, we observe that for newly formed ties, there exists a “U-shape” between tie range and persistence probability; newly formed long ties have the highest persistence probability (see Supplementary Note 7). This finding contradicts our conjecture that the persistence probability of newly formed long ties would be the smallest. Thus, for the two notions we examined, we find that (1) is supported while (2) is not supported. Therefore, the survival bias hypothesis does not fully explain our main results.

Valuable long ties. Our last hypothesis is that long ties tend to be more valuable. This hypothesis is consistent with weak tie theory and the roles of long ties, as conjectured in previous studies^{1,14}. However, while most computational models that simulate real-world networks highlight homophily⁴³—the phenomenon that individuals with similar attributes tend to be friends—previous models do not typically consider the benefits of social exchange

between people with different skills or information sets⁴². Recent work⁴², provides an example of how one can consider homophily and social exchange jointly, but this work is restricted to static social networks. Below, we propose a computational model that combines game theory and machine learning in order to examine long tie dynamics. This model helps support our hypothesis on valuable long ties, while also incorporating the first two hypotheses.

The model explaining long ties’ persistency. Here, we propose a game-theoretical computational model that simulates the dynamics of social networks. Specifically, the model combines the embedding techniques in machine learning^{39–41,44} and the strategic network formation in economics^{7,45}. Compared to the common network formation game models in the economics literature, our model stresses the high-dimensional heterogeneity, as well as the values of social exchange. Compared to network embedding techniques, our model helps understand the social network formation mechanisms. Ultimately, our model integrates the strategic network formation approach to explain the mechanisms, while the embedding techniques improve the

predictability of the computational model. Our study echoes Hofman’s recent paper that discusses the trade-off between explanation and prediction in computational social science⁴⁶.

Our model considers two procedures during the formation of social ties: the meeting procedure, and the choice procedure. This two-step model takes into account the dynamics of social ties – that people first meet others randomly, and then make their rational decisions about the choice of friends. The meeting procedure models reality, wherein people meet each other at random. There may exist many potential neighbor candidates who are mutually beneficial (e.g., some potentially valuable long ties), but the extremely low meeting probability can prevent the social tie from being formed. Moreover, when first meeting a new neighbor, a person may lack sufficient information to assess the person, and they are unable to make a rational decision about the social tie. After getting to know a new friend over a period of time (one phase in our study), the individual can then start to make a rational decision about that person. The choice procedure assumes that individuals are rational when choosing their network neighbors and that each individual maximizes their utility function.

Formally, let \mathcal{I} be the set of individuals and let i (or j, ℓ) be their index. Additionally, let t index the discrete time steps (or phases), and thus, $t \in \mathbb{N}^+$. Also, let $\mathbf{A}^{(t)}$ denote the adjacency matrix in phase t . $\mathbf{A}_{ij}^{(t)} = 1$ indicates that i and j are connected in phase t . $\mathbf{A}_{ij}^{(t)} = 0$ indicates that i and j are disconnected in phase t . For simplicity, we only consider an undirected network, i.e., $\mathbf{A}_{ij}^{(t)} = \mathbf{A}_{ji}^{(t)}$ for all $i, j \in \mathcal{I}$, and for all, $t \in \mathbb{N}^+$. To account for the heterogeneity of individual attributes, we use the “endowment vector” \mathbf{w}_i , which is a K -dimensional vector as in the embedding techniques^{39,40}. As embedding techniques do, each dimension measures a certain latent attribute of an individual, such as a type of skill or useful information. A larger w_{ik} indicates that the individual retains a high endowment of the k^{th} dimension.

In each phase, the neighbor’s set of i consists of two components: the new friend set $\mathcal{M}_i^{(t)}$, and the existing friend set $\mathcal{N}_i^{(t)}$; which echoes our analysis newly formed ties and existing ties. The new friend set is formed in the random meeting procedure. We assume each pair of individuals has a different meeting probability. The concept of a “meeting probability” is found widely in several econometric studies that aim to model social network formation^{45,47,48}. Specifically, for each pair of individuals, i and j , they have a probability of $p_{ij}^{(t)}$ to “meet” each other in phase t . If $\mathbf{A}_{ij}^{(t-1)} = 1$, that is, the two individuals were connected in phase $t - 1$, then the $p_{ij}^{(t)}$ is a large probability. Otherwise, $p_{ij}^{(t)}$ is a small probability, dependent on the network topology between i and j . Inspired by our previous comparison between newly formed ties and existing ties, we can imagine that if this is a long tie, the probability would be much smaller. Formally, we parametrize $p_{ij}^{(t)}$ as follows:

$$p_{ij}^{(t)} = \begin{cases} d_{t-1}(i, j) & \mathbf{A}_{ij}^{(t-1)} = 0 \\ q & \mathbf{A}_{ij}^{(t-1)} = 1 \end{cases} \quad (2)$$

The distance metric $d_{t-1}(i, j)$ depends on the network topology between individual i and individual j in phase $t - 1$. We define the distance metric to be proportional to the probability of random walks from i to j . Here, q is set to describe the probability of maintaining the meeting procedure in phase t .

The second component is the existing friend set $\mathcal{N}_i^{(t)}$, which is determined by the rational choice procedure. It is a subset of all friends in phase $t - 1$, i.e., $\mathcal{N}_i^{(t)} \in \mathcal{M}_i^{(t-1)} \cup \mathcal{N}_i^{(t-1)}$. This means

that individuals make rational decisions after maintaining their friendships for a period of one phase. The rationale behind this notion is that individuals need a significant amount of time to assess the value of an existing friend, so the rational choice procedure happens in the phase immediately following the meeting procedure. For a connected social tie in phase $t - 1$, the friendship must survive both the meeting procedure (a random draw from $\text{Bern}(q)$) and the rational choice procedure. The choice procedure is modeled using the following utility function:

$$U_i^{(t)}(c_{ij}^{(t)}) = \sum_{j \in \mathcal{M}_i^{(t-1)} \cup \mathcal{N}_i^{(t-1)}} \left(c_{ij}^{(t)} \sum_k \left(\sigma(w_{jk} - w_{ik}) + \sum_{\ell \in \mathcal{M}_j^{(t-1)} \cup \mathcal{N}_j^{(t-1)}} \delta \sigma(w_{\ell k} - w_{ik}) \right) - \left(c_{ij}^{(t)} \right)^2 \right), \text{ where } \sum_j \left(c_{ij}^{(t)} \right)^2 = 1. \quad (3)$$

Here, $U_i^{(t)}$ is the utility function of individual i in phase t . $c_{ij}^{(t)} \in [0, 1]^{\mathcal{M}_i^{(t-1)} \cup \mathcal{N}_i^{(t-1)}}$, which can be understood as a function that maps any j in the neighbor set in phase $t - 1$, i.e., each element in $\mathcal{M}_i^{(t-1)} \cup \mathcal{N}_i^{(t-1)}$, to a real number in $[0, 1]$. The utility function sums over all i ’s neighbors in phase $t - 1$. σ is the ReLU function: if $w_{jk} - w_{ik} > 0$, the output is $w_{jk} - w_{ik}$; otherwise, 0. ℓ enumerates over all j ’s neighbors in phase $t - 1$, which are also i ’s “friends’ friends.” The depreciation factor δ , which ranges in $(0, 1)$, measures how the value of a potential friend depreciates as the distance on the network increases. We refer to $\sigma(w_{jk} - w_{ik}) + \sum_{\ell \in \mathcal{M}_j^{(t-1)} \cup \mathcal{N}_j^{(t-1)}} \delta \sigma(w_{\ell k} - w_{ik})$ as the benefit that j brings to i . In addition, we separate the benefit into two: the direct benefit, $\sigma(w_{jk} - w_{ik})$, and the indirect benefit, $\sum_{\ell \in \mathcal{M}_j^{(t-1)} \cup \mathcal{N}_j^{(t-1)}} \delta \sigma(w_{\ell k} - w_{ik})$. The design of these benefit terms was intended for our valuable long tie hypothesis – we hope to observe that long ties have, on average, larger values in the direct benefit term.

$c_{ij}^{(t)}$ measures the time investment of i in j . A non-zero value of $c_{ij}^{(t)}$ indicates that j belongs to $\mathcal{N}_i^{(t)}$. The restriction of the sum of squared $c_{ij}^{(t)}$ reflects that people have limited time or energy to invest in their neighbors. The benefit of each neighbor is proportional to the time or energy investment in each neighbor j ; this is why we multiply the benefit term by $c_{ij}^{(t)}$. At the same time, the squared term $\left(c_{ij}^{(t)} \right)^2$ is used to measure the cost of time or energy. The design of $c_{ij}^{(t)}$ echoes our degree heterogeneity hypothesis – that those with many ties may have less investment in any one individual neighbor.

By the Cauchy-Schwarz inequality, Eq. (3) can be solved by

$$\left(c_{ij}^{(t)} \right)^* \propto \sum_k \left(\sigma(w_{jk} - w_{ik}) + \sum_{\ell \in \mathcal{M}_j^{(t-1)} \cup \mathcal{N}_j^{(t-1)}} \delta \sigma(w_{\ell k} - w_{ik}) \right), \text{ and } \sum_j \left(\left(c_{ij}^{(t)} \right)^* \right)^2 = 1. \quad (4)$$

In particular,

$$j \in \mathcal{N}_i^{(t)} \text{ iff } \left(c_{ij}^{(t)} \right)^* > 0; j \notin \mathcal{N}_i^{(t)} \text{ iff } \left(c_{ij}^{(t)} \right)^* = 0. \quad (5)$$

In other words, if the optimal solution informs $\left(c_{ij}^{(t)} \right)^* = 0$, then this indicates that i and j are no longer connected. Otherwise, $\left(c_{ij}^{(t)} \right)^*$ is the fraction of the call duration during which i interacts with j at time t among i ’s total call duration at time t .

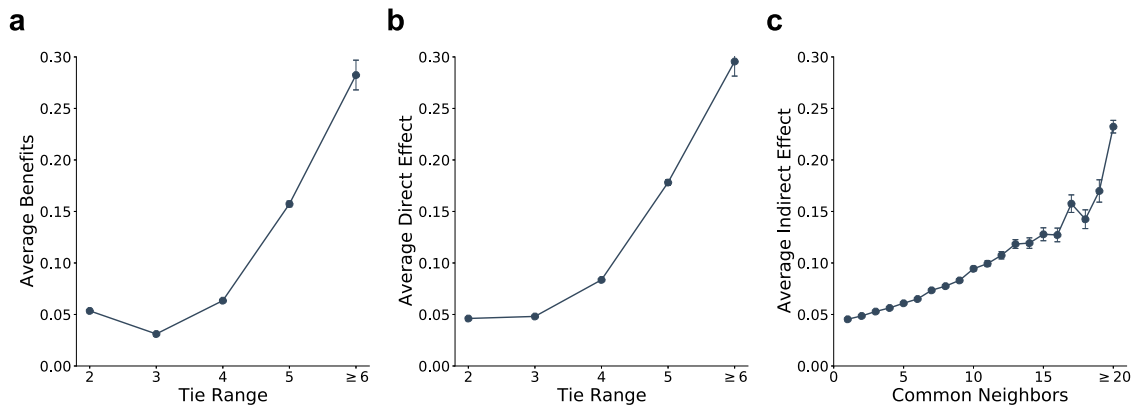


Fig. 7 Average benefits, average direct effects, and average indirect effects learned by our model. The curves represent (a) the average benefits, (b) average direct effects, and (c) average indirect effects for each tie range learned from our model, respectively. Error bars are 95% confidence intervals for the benefits (assuming normal distribution). Note that error bars are sometimes smaller than the data point markers.

This model provides major improvements based on the framework proposed in prior work⁴². First, different from their paper, we establish a model for network dynamics. In particular, we incorporate a meeting procedure; this addresses the phenomenon that, in reality, there are many neighbor candidates who do not form links purely because they have no opportunity to meet. Second, our model also takes into account the “weight” (i.e., the interaction frequency or duration) of the links. This is different from Yuan et al.⁴², where the weights between the links are binary. Third, Yuan et al.⁴² assumes that the marginal utility of additional neighbors is not dependent on other existing neighbors; by contrast, our model does not incorporate this assumption, and it also accounts for the network externality (i.e., the benefits of friends of friends)⁷. We provide additional analyses to verify our modeling fitting capacity in Supplementary Note 8.

Figure 7 provides the main implications derived from the learning results of our model. We first present the average benefit, i.e., $\sigma(w_{jk} - w_{ik}) + \sum_{\ell \in \mathcal{M}_j^{(t-1)} \cup \mathcal{N}_j^{(t-1)}} \delta\sigma(w_{\ell k} - w_{ik})$, given the different tie range in Panel (a) of Fig. 7. The average is taken over all candidate neighbors in $\mathcal{M}_j^{(t-1)} \cup \mathcal{N}_j^{(t-1)}$ given the tie range in phase $t - 1$. From this, we find a “U-shape”, i.e., the average benefit decreases with the tie range at the beginning, but later increases with the tie range. This is consistent with our previous findings regarding the “U-shape” between tie range and tie strength.

Next, we separate the benefits in Eq. (3) into the direct effect and the indirect effect. We present the average direct effect, which is $\sigma(w_{jk} - w_{ik})$ in Panel (b) of Fig. 7. We observe an increasing pattern with the tie range, indicating that as the tie range increases, the average benefit that a tie brings also increases. This result supports our hypothesis that long ties tend to be more valuable, which also explains the results in the previous sections. We also compute the average indirect effect, i.e., $\sum_{\ell \in \mathcal{M}_j^{(t-1)} \cup \mathcal{N}_j^{(t-1)}} \delta\sigma(w_{\ell k} - w_{ik})$. In our model, only social ties with common friends, i.e., those with a tie range of 2, have indirect effects. We plot the relationship between the number of common neighbors and the average indirect effect. The indirect effect echoes our previous discussion on patterns of social ties with a tie range of 2. As observed in Panel (c) of Fig. 7, we find an increasing pattern. In particular, by examining the first several data points in the plot, we observe a seemingly convex pattern, indicating the increasing marginal utility of common neighbors.

Overall, the results from our learning model suggest that long ties are generally more valuable (with greater direct effects). This

model also takes into account degree heterogeneity and survival bias hypotheses, although they are probably not the primary drivers. We also compare our model with other baseline models in Supplementary Note 9, but they cannot provide the implications as we plot in Fig. 7.

Conclusion

In this study, we combine empirical analysis and an interdisciplinary computational model to investigate the dynamics of long ties. We find that long ties persist longer than shorter-range ties and that many long ties are persistently long. These results are contrary to what is suggested by several prior theories and prediction models. To better understand our results, we propose three hypotheses—degree heterogeneity, survival bias, and valuable long ties—and then go on to discuss the limitations of both the degree heterogeneity hypothesis and the survival bias hypothesis. Finally, we discuss an interdisciplinary model that combines game theory and machine learning to support our valuable long-tie hypothesis. Verified by real-world data, our model partly explains why long ties are more persistent than what has previously been suggested by existing theories and models.

Our results also signal the importance of social interventions that promote the formation of long ties, such as mixing diverse people with diverse backgrounds. For example, both our empirical analysis and modeling results indicate that people who are dissimilar in certain attributes or who are distant in a social network may have significant mutual benefits to one another. However, as indicated by our model, the small likelihood of those people meeting can hinder the formation of their future interactions.

Based on this study, there are several interesting research directions that could be investigated. First, although we examine a large-scale social network with very few missing nodes, the generalizability of our results should be interrupted cautiously. On the one hand, our study replicates the U-shape in Park et al.²⁵ which examines multiple static phone communication and Twitter networks. The successful replication provides confidence in the potential generalizability of our additional dynamic analyses to these networks. On the other hand, there are many other types of social ties rather than phone communications, such as social media, offline interactions, or collaboration networks. We appeal for more studies on this important topic to verify the external validity of our conclusions. Second, although most existing studies on long ties, including ours, use the aggregate data to measure the tie range, it is interesting to investigate how

to leverage advanced methods of analyzing temporal networks to further understand the mechanisms of dynamics of long ties, which can examine events occurring on network paths on a more fine-grained level^{49–51}. Finally, there may be intriguing variants of our model. For example, our model only reflects the absolute advantages that other people bring, but it would be interesting to incorporate comparative advantages in our model, as well.

Methods

Data description. In our study, we use a nationwide call detail record dataset. Users’ private information has been anonymized and thus we are unable to identify them. This data provider is a company that functions as the main service provider for most of the mobile phone users in a European region. The time period covered by the data starts from Jan. 2015 to Dec. 2016. In the dataset, we retrieve the total number of calls, texts, as well as the duration of calls between any two people in each month. See Supplementary Note 1 for more details.

We establish a temporal social network with the dataset. We consider discrete time steps (or phases): for each phase, we construct a “snapshot” of the network, where the node indicates a user and the edge represents the interaction between two users. A key question is how we determine the length of the time window of each phase. In our main results, we treat every three months as a phase. In Supplementary Note 2, we also use one month or six months to verify the robustness of our results.

To maintain a temporal network where the node set is stable and the global network structure does not change dramatically with the dynamics of a few nodes, we only consider the interactions among users who have at least one call or text in every phase. We construct a temporal directed network with 45,192 nodes and 385,533 edges on average for each phase.

In terms of the weight of the directed network, we consider two variables as mentioned in the main text: interaction frequency and duration. Interaction frequency is the total number of calls or text that node i sends to j ; there are a few calls with zero-second duration and we filter those calls out. Interaction duration is the total time length that i calls j in each phase, and does not account for texting.

Tie range and long ties. Tie range^{14,25} is defined as the length of the second shortest path between two connected nodes (Fig. 1). It indirectly reflects the network distance of the connection. Consistent with previous long tie studies^{22,25}, there is no clear cutoff of tie range that decides whether a tie is short or long. A good reference is the Milgram experiment, which suggested that the average network distance between every two people is ~6. In our study, we treat social ties with a tie range of 2 as short ties, and ties with 5 or ≥6 as long ties. Besides, we do a sensitive check of our results by randomly dropping a proportion (5%) of nodes or edges (see Supplementary Note 3). Our main results are verified not sensitive to a few nodes or edges happening to exist on the network.

Details in learning. Based on Eq. (4), we construct the loss function to minimize the MSE Loss between c_{ij} and its right hand side. We use stochastic gradient descent to optimize the loss function. For each epoch, we construct our loss function as below:

$$\mathcal{L} = \mathcal{L}_{pos} + \mathcal{L}_{neg}, \tag{6}$$

The loss function is composed of the loss functions of positive (connected pairs), and negative samples (disconnected pairs).

$$\mathcal{L}_{pos} = \sum_{i \in \text{sampled}} \left(\frac{\sum_{j \in (\mathcal{N}_i^{(t-1)} \cup \mathcal{M}_i^{(t-1)}) \cap \mathcal{N}_i^{(t)}} |c_{ij}^{(t)} - \hat{c}_{ij}^{(t)}|}{\sum_{j \in (\mathcal{N}_i^{(t-1)} \cup \mathcal{M}_i^{(t-1)}) \cap \mathcal{N}_i^{(t)}} 1} \right); \tag{7}$$

$$\mathcal{L}_{neg} = \sum_{i \in \text{sampled}} \left(\frac{\sum_{j \in (\mathcal{N}_i^{(t-1)} \cup \mathcal{M}_i^{(t-1)}) \cap \mathcal{N}_i^{(t)}} \hat{c}_{ij}^{(t)}}{\sum_{j \in (\mathcal{N}_i^{(t-1)} \cup \mathcal{M}_i^{(t-1)}) \cap \mathcal{N}_i^{(t)}} 1} \right). \tag{8}$$

The set “sampled” denotes the set of sampled nodes in each epoch. For positive samples, we minimize the difference between $c_{ij}^{(t)}$, the time investment of i on j , and the predicted time investment denoted by $\hat{c}_{ij}^{(t)}$.

$$\hat{c}_{ij}^{(t)} = \frac{\log(D_{ij}^{(t)} + 1)}{\sum_{j \in \mathcal{M}_i^{(t-1)} \cup \mathcal{N}_i^{(t-1)}} \log(D_{ij}^{(t)} + 1)}, \tag{9}$$

where $D_{ij}^{(t)}$ is the interaction duration between i and j in phase t . To reduce the

impact of extreme values, we take the logarithm of $D_{ij}^{(t)}$. Since $D_{ij}^{(t)} \geq 0$, $\hat{c}_{ij}^{(t)} \geq 0$.

$$\hat{c}_{ij}^{(t)} = \frac{\exp \left\{ \sum_k \left(\sigma(w_{jk} - w_{ik}) + \sum_{\ell \in \mathcal{M}_j^{(t-1)} \cup \mathcal{N}_j^{(t-1)}} \delta \sigma(w_{\ell k} - w_{ik}) \right) \right\}}{\sum_{j \in \mathcal{M}_i^{(t-1)} \cup \mathcal{N}_i^{(t-1)}} \exp \left\{ \sum_k \left(\sigma(w_{jk} - w_{ik}) + \sum_{\ell \in \mathcal{M}_j^{(t-1)} \cup \mathcal{N}_j^{(t-1)}} \delta \sigma(w_{\ell k} - w_{ik}) \right) \right\}}. \tag{10}$$

When minimizing the loss function, we treat the time investment of i in j , which is calculated by the interaction duration or frequency, as the input and endowment vectors in this loss function as the variables to be inferred. Note that the existence of the δ may result in an uncontrollable gradient issue. We thus use grid search for this variable and check the robustness of our results in Supplementary Note 8. Moreover, we also discuss the selection of the number of dimensions of the endowment vectors in Supplementary Note 8.

To facilitate the learning process, we apply mini-batch stochastic gradient descent with Adam optimizer⁵². Consistent with conventional network embedding algorithms, node sampling probability is proportional to node degree (d_i^2)⁵³. In this case, the endowment vectors of both these sampled nodes and their neighbors will be updated in each epoch in the gradient descent. In Supplementary Note 8, we show that our learning converges under this setting. Details in the machine learning implementation are also discussed in Supplementary Note 8.

Data availability

Data is available at <https://github.com/DingLyu/Investigating-and-Modeling-the-Dynamics-of-Long-Ties>. Differential privacy is applied to protect the privacy of users.

Code availability

Code is available at <https://github.com/DingLyu/Investigating-and-Modeling-the-Dynamics-of-Long-Ties>.

Received: 23 September 2021; Accepted: 17 March 2022;

Published online: 11 April 2022

References

- Watts, D. J. & Strogatz, S. H. Collective dynamics of ‘small-world’ networks. *Nature* **393**, 440 (1998).
- Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
- Jackson, M. O. *Social and Economic Networks*. (Princeton Univ. Press, Princeton, 2010).
- Barabási, A.-L. *Network Science*. (Cambridge Univ. Press, Cambridge, 2016).
- Broido, A. D. & Clauset, A. Scale-free networks are rare. *Nat. Commun.* **10**, 1–10 (2019).
- McPherson, J. M., Popielarz, P. A. & Drobnic, S. Social networks and organizational dynamics. *Am. Sociol. Rev.* **57**, 153–170 (1992).
- Jackson, M. O. & Wolinsky, A. A strategic model of social and economic networks. *J. Econ. Theory* **71**, 44–74 (1996).
- Clauset, A., Newman, M. E. & Moore, C. Finding community structure in very large networks. *Phys. Rev. E* **70**, 066111 (2004).
- Liben-Nowell, D. & Kleinberg, J. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.* **58**, 1019–1031 (2007).
- Christakis, N. A. & Fowler, J. H. The spread of obesity in a large social network over 32 years. *N. Engl. J. Med.* **357**, 370–379 (2007).
- Entwisle, B., Faust, K., Rindfuss, R. R. & Kaneda, T. Networks and contexts: Variation in the structure of social ties. *Am. J. Sociol.* **112**, 1495–1533 (2007).
- Flache, A. & Macy, M. W. The weakness of strong ties: collective action failure in a highly cohesive group. In *Evolution of Social Networks*, 27–52 (Routledge, 2013).
- Burt, R. S. *Structural Holes*. (Harvard Univ. Press, Cambridge, 1992).
- Granovetter, M. S. The strength of weak ties. *Am. J. Sociol.* **78**, 1360–1380 (1973).
- Levin, D. Z. & Cross, R. The strength of weak ties you can trust: the mediating role of trust in effective knowledge transfer. *Manage. Sci.* **50**, 1477–1490 (2004).
- Onnela, J.-P. et al. Structure and tie strengths in mobile communication networks. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7332–7336 (2007).
- Zhao, J., Wu, J. & Xu, K. Weak ties: Subtle role of information diffusion in online social networks. *Phys. Rev. E* **82**, 016105 (2010).
- Ghasemiesfeh, G., Ebrahimi, R. & Gao, J. Complex contagion and the weakness of long ties in social networks: revisited. In *Proceedings of the ACM Conference on Electronic Commerce*, 507–524 (2013).
- Larson, J. M. The weakness of weak ties for novel information diffusion. *Appl. Netw. Sci.* **2**, 1–15 (2017).

20. Gee, L. K., Jones, J. J., Fariss, C. J., Burke, M. & Fowler, J. H. The paradox of weak ties in 55 countries. *J. Econ. Behav. Organ.* **133**, 362–372 (2017).
21. Montgomery, J. D. Weak ties, employment, and inequality: an equilibrium analysis. *Am. J. Sociol.* **99**, 1212–1236 (1994).
22. Centola, D. & Macy, M. Complex contagions and the weakness of long ties. *Am. J. Sociol.* **113**, 702–734 (2007).
23. Centola, D. The spread of behavior in an online social network experiment. *Science* **329**, 1194–1197 (2010).
24. Romero, D. M., Meeder, B. & Kleinberg, J. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of International Conference on World Wide Web*, 695–704 (2011).
25. Park, P. S., Blumenstock, J. E. & Macy, M. W. The strength of long-range ties in population-scale social networks. *Science* **362**, 1410–1413 (2018).
26. Trieu, P., Bayer, J. B., Ellison, N. B., Schoenebeck, S. & Falk, E. Who likes to be reachable? availability preferences, weak ties, and bridging social capital. *Inform. Commun. Soc.* **22**, 1096–1111 (2019).
27. Aral, S. & Van Alstyne, M. The diversity-bandwidth trade-off. *Am. J. Sociol.* **117**, 90–171 (2011).
28. Todo, Y., Matous, P. & Inoue, H. The strength of long ties and the weakness of strong ties: Knowledge diffusion through supply chain networks. *Res. Policy* **45**, 1890–1906 (2016).
29. Eckles, D., Mossel, E., Rahimian, M. A. & Sen, S. Long ties accelerate noisy threshold-based contagions. *Preprint at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3262749* (2019).
30. Jahani, E., Fraiberger, S., Bailey, M. & Eckles, D. Origins and consequences of long ties in social networks. *Preprint at <https://osf.io/preprints/socarxiv/g2nkq/>* (2022).
31. Block, P. Reciprocity, transitivity, and the mysterious three-cycle. *Soc. Netw.* **40**, 163–173 (2015).
32. Li, A., Cornelius, S. P., Liu, Y.-Y., Wang, L. & Barabási, A.-L. The fundamental advantages of temporal networks. *Science* **358**, 1042–1046 (2017).
33. Easley, D. et al. *Networks, Crowds, and Markets*. (Cambridge univ. press, Cambridge, 2010).
34. Benson, A. R., Abebe, R., Schaub, M. T., Jadbabaie, A. & Kleinberg, J. Simplicial closure and higher-order link prediction. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E11221–E11230 (2018).
35. Asikainen, A., Iñiguez, G., Ureña-Carrión, J., Kaski, K. & Kivela, M. Cumulative effects of triadic closure and homophily in social networks. *Sci. Adv.* **6**, eaax7310 (2020).
36. Brashears, M. E. & Quintane, E. The weakness of tie strength. *Soc. Netw.* **55**, 104–115 (2018).
37. Santos, F. C., Pacheco, J. M. & Lenaerts, T. Cooperation prevails when individuals adjust their social ties. *PLoS Comput. Biol.* **2**, e140 (2006).
38. Weng, L., Karsai, M., Perra, N., Menczer, F. & Flammini, A. Attention on weak ties in social and communication networks. In *Complex Spreading Phenomena in Social Systems*, 213–228 (Springer, 2018).
39. Perozzi, B., Al-Rfou, R. & Skiena, S. Deepwalk: Online learning of social representations. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 701–710 (2014).
40. Grover, A. & Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 855–864 (2016).
41. Veličković, P. et al. Graph Attention Networks. In *Proceedings of the International Conference on Learning Representations* (2018).
42. Yuan, Y., Alabdulkareem, A. & Pentland, A. S. An interpretable approach for social network formation among heterogeneous agents. *Nat. Commun.* **9**, 1–9 (2018).
43. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.* **27**, 415–444 (2001).
44. Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations* (2017).
45. Christakis, N., Fowler, J., Imbens, G. W. & Kalyanaraman, K. An empirical model for strategic network formation. In *The Econometric Analysis of Network Data*, 123–148 (Elsevier, 2020).
46. Hofman, J. M. et al. Integrating explanation and prediction in computational social science. *Nature* **595**, 181–188 (2021).
47. Mele, A. A structural model of dense network formation. *Econometrica* **85**, 825–850 (2017).
48. Overgoor, J., Benson, A. & Ugander, J. Choosing to grow a graph: modeling network formation as discrete choice. In *Proceedings of the International Conference on World Wide Web*, 1409–1420 (2019).
49. Holme, P. & Saramäki, J. Temporal networks. *Phys. Rep.* **519**, 97–125 (2012).
50. Holme, P. Modern temporal network theory: a colloquium. *Eur. Phys. J. B* **88**, 1–30 (2015).
51. Sekara, V., Stopczynski, A. & Lehmann, S. Fundamental structures of dynamic social networks. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 9977–9982 (2016).
52. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations* (2015).
53. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S. & Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, 3111–3119 (2013).

Acknowledgements

The authors are grateful for the comments and suggestions made by three anonymous reviewers and the editors.

Author contributions

D.L. and Y.Y. conceived the present idea. Y.Y. collected and processed the data. D.L. and Y.Y. analyzed the results. D.L., Y.Y., L.W., X.W., and A.P. discussed the analytical approach and furthered the results. D.L. and Y.Y. wrote the paper with input from L.W., X.W., and A.P. All authors have reviewed and commented on the manuscript.

Competing interests

The authors declare no competing interests.

Ethical declaration

Our study has been determined to be exempt by MIT IRB (COUHES). Exempt ID: E-3442.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s42005-022-00863-w>.

Correspondence and requests for materials should be addressed to Yuan Yuan.

Peer review information *Communications Physics* thanks Gerardo Iñiguez, Christian Steglich and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022