On the Friendship Paradox and Inversity: A Network Property with Applications to Privacy-sensitive Network Interventions

Vineet Kumar[a,2], **David Krackhardt**[b], **and Scott Feld**[c]

This manuscript was compiled on July 30, 2023

We provide the mathematical and empirical foundations of the friendship paradox (FP) in networks, often stated: "Your friends have more friends than you." We prove a set of network properties on friends of friends, and characterize the concepts of local and global means. We propose a network property called Inversity, which quantifies the imbalance in degrees across edges, and prove that the sign of Inversity determines the ordering between local or global means for any network, with implications for interventions. Network intervention problems benefit from using highly-connected nodes, e.g. immunization. We characterize two intervention strategies based on the friendship paradox to obtain such nodes, with the novel global and local strategy. Both strategies are informationally light and privacy-sensitive, and do not require the network structure to provide provably guaranteed improvements. We demonstrate that the proposed strategies obtain several-fold improvement (100-fold in some networks) in node degree relative to a random benchmark, for both generated and real networks. We evaluate how Inversity informs which strategy would work better based on network structure, and show how network aggregation can alter Inversity. We illustrate how these strategies can be used to control contagion of an epidemic spreading across a set of village networks, finding that these strategies require far fewer nodes to be immunized (less than 50% relative to random). The interventions do not require knowledge of network structure, are privacy-sensitive, flexible for time-sensitive action, only requiring selected nodes to nominate network neighbors for intervention.

Network Intervention | Friendship Paradox | Inversity | Contagion

Significance Statement

Networks across many different settings, including social, economic and natural, are powerful tools for interventions due to the cascading impact of one individual on others. All networks with degree variation exhibit the friendship paradox phenomenon. We demonstrate its multifaceted nature, and provide its foundations mathematically and empirically. and identify a new network property called Inversity. The proposed network intervention strategies provide a privacy-sensitive approach to obtaining highly connected individuals, without knowing the network. These strategies are guaranteed to obtain (weakly) greater than average degree for any network, and are useful in a variety of applications. We characterize the value of these strategies theoretically with real-world networks. Finally, we show how to identify which strategy performs best using Inversity.

We examine the underlying mathematical and empirical foundations of the friendship paradox, and define a new network property called Inversity, which has implications for network interventions. The friendship paradox has often been simply referred to by the maxim "Your friends have more friends than you do." However, we show that there are two different ways of understanding this statement, which lead to different network properties that we call local and global mean number of friends of friends. We find that both means are higher than the average degree across nodes in the network. We show that the properties are not just conceptually distinct, but they are also empirically different across a wide class of generated and real-world networks. We identify a novel network property, Inversity, that connects the two means, and for any network, the sign of Inversity determines whether the local mean or global mean is higher.

The above results have direct implications for interventions by finding highly connected nodes in a network using privacy-sensitive methods based on the friendship paradox. The two means lead to corresponding local and global strategies for obtaining highly connected nodes, of which the global strategy has not been used in network interventions. The Inversity of the network indicates which strategy (local or global) is better for identifying highly connected nodes. We show in empirical real-world networks and in generated networks that the strategy used can make a meaningful difference. We show in a simplified application using real networks that using inversity to choose the local or global strategy to identify inoculation candidates would considerably reduce the epidemic threshold and peak infection relative to random selection.

Friendship Paradox

The Friendship Paradox, which our interventions are based on, is colloquially stated as the idea that people's friends

Author affiliations: [a]Yale School of Management, Yale University, 165 Whitney Avenue, New Haven CT 06511. ORCID: 0000-0001-8784-6858; [b]Heinz College, Carnegie Mellon University, 4800 Forbes Avenue, Pittsburgh, PA 15213. ORCID: 0000-0001-9487-9973; [c]Department of Sociology, College of Liberal Arts, Purdue University, 700 West State Street, West Lafayette, IN 47907-2059. ORCID: 0000-0003-4820-1365

Contributions

The authors have no competing interests as it relates to this paper.

[2]To whom correspondence should be addressed. E-mail: vineet.kumar@yale.edu

are more popular than them (1, 2).[*] The intuition for why the friendship paradox helps obtain well-connected nodes is this: since highly-connected nodes (hubs) are connected to many other nodes (by definition), obtaining a random friend (or neighbor) of a random node is likely to result in hubs with greater likelihood, compared to the case of randomly selecting nodes.

We establish that the friendship paradox is not just one statement, but a set of distinct claims (All theorems and proofs are in Supplement §S.B). First, we find an impossibility, i.e. the individual-level friendship paradox cannot hold for all individuals in any given network (Theorem S1). In practice for real networks, it can hold for a large proportion of nodes in the network (Figs. S3 and S4 in Supplement §S.D). Second, we demonstrate that in contrast to the impossibility of the individual friendship paradox, the network level friendship paradox holds for any network. We find that the *average* number of friends of friends across the network can be characterized in two different ways, using the local and global mean, defined below. Both local and global means are greater than the mean degree of the network, and these means are related through a novel network characteristic we term **Inversity**.

Local and Global Means. We formally characterize the two distinct but related network properties deriving from the friendship paradox relating to the "average number of friends of friends." Denote a network (see Table S1 for full notation) as an undirected graph $\mathcal{G} = (V, E)$ with V the set of nodes and E the set of edges ($e_{ij} \in \{0, 1\}$ denoting absence or presence of a connection between i and j), D_i refers to the degree of node i, and $\mathcal{N}(i)$ the set of i's neighbors. We specify the local mean as:

$$\mu_L = \frac{1}{N} \sum_{i \in V} \left[\frac{1}{D_i} \sum_{j \in \mathcal{N}(i)} D_j \right] \qquad [1]$$

The global mean is defined as the ratio of the total number of friends of friends to the total number of friends in the network, consistent with (1):

$$\mu_G = \frac{\sum_{i \in V} \left[\sum_{j \in \mathcal{N}(i)} D_j \right]}{\sum_{i \in V} D_i} \qquad [2]$$

The above means arise from differently weighting the average degree across friends. Both means above are consistent with the notion of "average number of friends of friends," although they are distinct network properties (see Fig. S1 for an example and detailed explanation). The global mean was theoretically investigated earlier and found to be greater than (or equal to) the average degree and is independent of the local structure of connections, given node degrees (Theorem S2). Equality holds only when the network is regular, with all nodes the same degree within and across components.[†]

The local mean is also shown to be greater than (or equal to) the mean degree (Theorem S3).[‡] However, the contrast is that the local mean has distinct properties that depend on local network structure (i.e. who is connected to whom). Equality for the local mean only holds when each component is regular, with no degree variation within components.

We identify network structures that result in a greater divergence between the Local and Global mean, and between these means and the average degree, including whether one of the means is always greater than the other, whether they always exhibit correlated variation away from the mean degree. In Figure S2, we find that both local and global mean can be much greater than the mean degree, and between these two means, either of them can be greater than the other, and in some network structures, both can be relatively high compared to the mean degree. We also see the global mean is invariant to rewiring the network while keeping the degree distribution the same, whereas the local mean is impacted by rewiring (Theorem S6).

Inversity: Connecting Local and Global Means

We identify and define a network property, Inversity, that determines when the local mean is greater than the global mean. This property captures all local network information related to the local mean and is scale-invariant, i.e. independent of the size or density of the network. We prove that the sign of inversity helps us determine which of the local mean or the global mean is higher for any given network. We show how inversity is related to but distinct from degree assortativity (in Supplement §S.G).

Inversity is a correlation-based metric that relates the global and local means for any network is obtained as follows. First, define the following edge-based distributions to examine the relationship between the means. The *origin* degree (**O**), $D^O(e)$, *destination* degree (**D**), $D^D(e)$, and *inverse destination* degree (**ID**) distribution, $D^{ID}(e)$, are defined across directed edges $e \in \hat{E}$ as: $D^O(e_{jk}) = D_j, D^D(e_{jk}) = D_k, D^{ID}(e_{jk}) = \frac{1}{D_k}$. We define the *inversity* across the edge distribution as the Pearson correlation across the origin and inverse degree distributions.

$$\rho = Corr\left(D^{\mathbf{O}}, D^{\mathbf{ID}}\right) \qquad [3]$$

We connect (see Theorem S4) the local and global means with *inversity* and the degree distribution $\left(\kappa_m = \sum_{i \in V} D_i^m\right)$ as:

$$\mu_L = \mu_G + \rho\ \Psi(\kappa_{-1}, \kappa_1, \kappa_2, \kappa_3) \qquad [4]$$

where Ψ is a positive function of the degree distribution.

When connections (edges) are mostly between nodes of similar degree, then inversity ρ is likely to be more negative. In such a case, the global mean is greater than the local mean. In contrast, when connections are more likely to be between nodes of dissimilar degree, then inversity ρ is positive, and the local strategy is likely to be obtain higher degree nodes.

Therefore, if inversity is known, we don't need the entire degree distribution to obtain the local mean. Rather, *four* moments of the degree distribution are sufficient for that purpose. Inversity captures the local information on imbalances

[*] The phenomenon has also been generalized to the idea that individual attributes and degree are correlated (3), e.g. an individual's co-authors are more likely to be cited (4), or that friends are more important (5), or more socially active (6). Of specific connection with this paper is the mathematical generalization to distributions examined in (7).

[†] There are a number of phenomena that share a similar underlying structure, e.g. disproportionately many people grow up in large families, or students experience of class size is higher than the average size of classes. The underlying selection process here is commonly termed probability proportional to size (PPS) (8, 9). In the case of the friendship paradox, we find that the Global mean has a direct mathematical connection to PPS (friends have disproportionately more friends). However, the Local mean operates through a different mechanism.

[‡] We term these means local or global since the former depends on the local structure (who is connected to whom), whereas the latter only depends on the global network properties (degree distribution).

in degree of nodes across edges, whereas the moments of the degree distribution represent global information about the network. Inversity ρ has a critical role in determining whether the local or global mean is larger for a network; specifically, $\rho < 0$ indicates the global mean is higher than the local mean, whereas $\rho > 0$ indicates the opposite. Thus, knowing inversity can help us determine which strategy to use. Even computing inversity is information-light, requiring only the $2k$ distribution, which represents the degrees of nodes at the termini of each edge, rather than the entire network (10).

How Inversity Depends on the Structure of Connections

Inversity is the only term that depends on the structure of connections (who is connected to whom) in the relationship between local and global means. We examine how Inversity changes as we change the structure, while *simultaneously* preserving the degree distribution. In Figure 1, we start with a network with the minimum inversity $\rho = -1$, and then use the rewiring theorem (Theorem S6) to examine how Inversity increases while the degree distribution, and consequently, the mean degree, variance of degree, minimum and maximum degree, as well as the global mean, all remain fixed and identical across each of the networks (a)–(i). Specifically, each network has $N = 25$ nodes, $\mu_D = 3.5$, $\sigma_D^2 = 3.4$, $D_{\min} = 1$ and $D_{\max} = 8$.

The motivation for keeping degree distribution fixed across the networks is that the global mean does not change. Specifically, noting from panel (a), the degree distribution includes 16 nodes with degree 1 (the dyads), and 9 nodes with degree 8 (clique or complete subgraph).

We note the rewiring patterns, beginning with (a), which displays a network with a fully connected complete component with 9 nodes, and 8 dyads. This network has the lowest possible inversity of $\rho = -1$, consistent with the idea that no edge connects nodes of different degree, which is essential for inversity to be greater than its minimum value. Observe, that the global mean $\mu_G = 6.7 > \mu_D = 3.5$ for this network. We use the rewiring theorem (Theorem S6) to increase inversity; this approach *connects* low degree nodes to high degree nodes, while *removing* connections between nodes of intermediate degree. The rewiring increases the variation in the degrees of the nodes connected by an edge, as the network transforms from (a) to (b) and in each further step. We observe that the nodes in each dyad break up their edge (which connects nodes of identical degree), and connect to nodes in the large component, which contains high degree nodes. We next observe a star-like structure form, beginning with panel (d). Finally, as the star-like structure expands, in panel (h) and (i), we find that Inversity has changed sign to become positive.

A few general observations are worth noting. First, we see that Inversity and local mean are highly sensitive to network structure (who is connected to whom), whereas the gobal mean is impacted only by the degree distribution, specifically its mean and variance. Second, we note that networks which display little or no variation among the node degrees connected by an edge have negative inversity, like in network (a). Third, we find a wide range of possible networks and Inversity levels and local mean for a fixed global mean, ranging from negative to positive Inversity. Finally, the degree distribution can constrain the range of Inversity. We next examine the implications of these findings for network interventions.

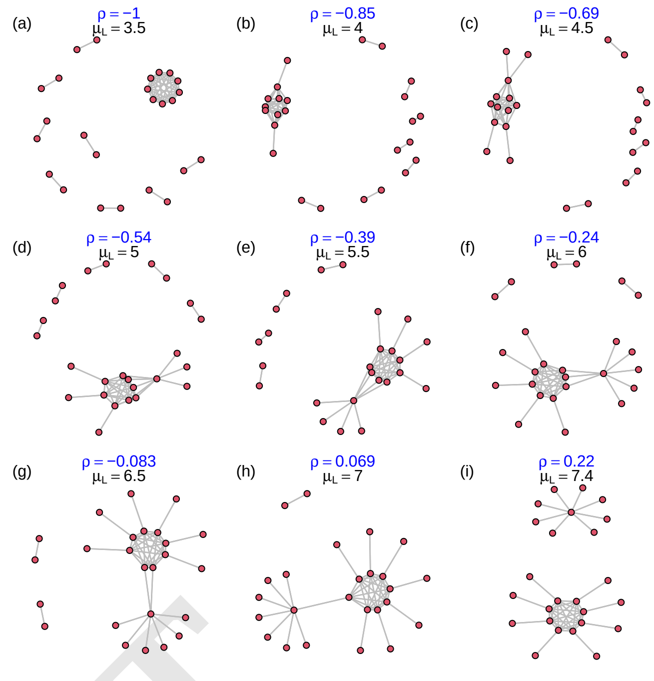


Fig. 1. How Inversity Changes with Rewiring. The network is changed by rewiring, starting with the top left ($\rho = -1$), to increase Inversity ρ as we traverse from panel (a) to (i). Observe that the number of nodes, $N = 25$, the number of edges $|E| = 44$, as well as the degree distribution for each of the networks in panels (a)–(i) is identical, with 16 nodes with degree 1 and 9 nodes with degree 8. We note that the local mean $\mu_L = \mu_D = 3.5$ for network (a), but increases along with Inversity in panels (b)–(i), reaching $\mu_L = 7.4$ for network (i). The Global mean $\mu_G = 6.7$ remains constant across all the networks.

There are many reasons why networks may take forms with high or low inversity. For the present purpose, we provide some intuition of relevant processes. We can expect hub-based (or star) networks to have high inversity, where most nodes have few ties that largely go to the relatively few hubs with large numbers of ties (e.g. Twitter celebrity based networks). In contrast, we can expect that networks of clusters of various sizes will have low inversity, where members of the large clusters tied to high degree and be tied to one another, and nodes in small clusters tend to have few ties and be tied to one another (e.g. friendship networks based upon group membership). The various causes of network structures having different levels of inversity may be the subject of extensive theoretical and empirical study in the future. We provide a discussion of this in Supplement §S.F. For details on inversity values in real-world networks and related findings, see Supplement §S.C.

Network Interventions

Consider the following: (a) (Reducing) A new infectious disease is spreading through a large population. We want to minimize the number of infected individuals by inoculating using a new vaccine; however, we only have a limited number of doses to administer. (b) (Accelerating) We have a new highly effective medical device with limited samples that we would like to provide to select medical professionals, who can then share information through word of mouth. (c) (Observing) We would like to identify viral contagion as quickly as possible by choosing individuals as observation stations (or for contact tracing). Although seemingly disparate, whether to reduce, accelerate or observe dynamic contagion, these problems represent a class of network interventions in which we benefit from identifying more central or highly connected individuals in the network (11).[§]

We show how seeding interventions using the friendship paradox local and global strategies developed here can impact network interventions, by helping to obtain highly connected seed nodes in a privacy-sensitive manner from the relevant network. Note that while the local strategy has been commonly suggested and used (14–16), its theoretical and empirical properties have not been examined and characterized for general networks. The global strategy is novel and first proposed here, and to our knowledge, has not been suggested or used for network interventions.

Our approach stands in contrast to most existing methods of identifying seeds for interventions, which focus on taking advantage of detailed network data on social connections, and even on activity to identify influential individuals (17–20). Privacy concerns are increasingly important in such settings, making it challenging to obtain network data (21–25). Users are also concerned that their data may be used in algorithms (26), and might even result in discrimination against them (27).

Relevant Network Structure: Even, if network data is available, the challenge in many cases is that we do not have access to the *relevant* network structure. In application (a), having the Facebook (or similar) network structure might not be useful, since the relevant network would be the *physical contact* network, which might be more challenging to obtain. In contrast, for application (b), finding a high degree node using a physical contact network of everyone who interacts with a medical professional is unlikely to be informative in characterizing opinion leadership in the profession. For (c), carrying out contact tracing for all individuals can be expensive in effort and time. These factors make it important to be able to leverage the structure of the relevant network while being sensitive to privacy concerns.

These friendship paradox based strategies have several advantages for implementation. First, despite being informationally-light, the strategies here provide provable advantages for virtually any network structures, in contrast to strategies that don't provide such guarantees for general networks. The network structure may be expensive to collect, not be possible to obtain in a timely manner, or may vary over time, making the proposed method more valuable. Second, the strategies are much more privacy-sensitive than mapping out social networks. Third, the strategies can be implemented quickly since they only require local network information obtained by querying individuals or interaction data. Finally, the class of interventions here can be used for both advance and consequent interventions, i.e. for both prevention and treatment interventions.

[§]We focus on the class of "simple contagion" problems, which require only one rather than "complex contagion" that require multiple exposures (12, 13).

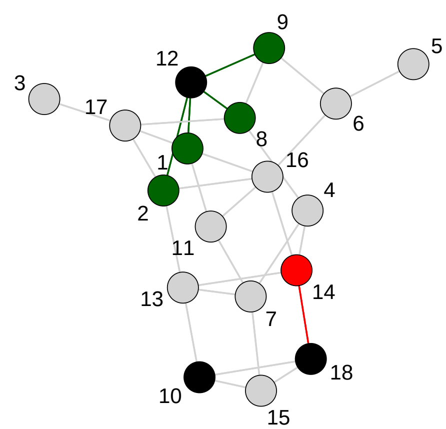
Implementing the Intervention Seeding Strategies

The above formulation of local and global mean suggests distinct strategies for choosing seeds for interventions, or intervention strategies. We illustrate *random*, *local* and *global* strategies to choose a "seed" node in the network beginning with an initial randomly chosen node (Table 1). The *local strategy* would query randomly selected individual nodes with the query, "could you suggest the name of a randomly chosen friend?" as example.

Table 1. Implementation of Seeding Strategies

Step	Details
0	Fix $p \in (0, 1]$ (only used for Global strategy in Step **2G**).
Repeat Steps 1-2 below until at least k seeds are present in the seed set $\mathcal{S}$.	
1	Draw a random node r uniformly from set of nodes, V. In Example Network, Nodes 10, 18 and 12 (in black) are drawn for (R), (L) and (G) strategies respectively.
2	Depending on the strategy Random (R), Local (L) or Global (G), do the following:
2R	(**Random**): Add r to the seed set $\mathcal{S}$. In Example Network, add node 10 to the seed set.
2L	(**Local**): Obtain a node s chosen with uniform probability from r's friends, i.e. $s \in \mathcal{N}_r$. Add the friend s to the seed set $\mathcal{S}$. In Example Network, one of node 18's friends, node 14 (in red), is chosen at random. Add node 14 to the seed set.
2G	(**Global**): For each of r's friends, $s \in \mathcal{N}_r$: With probability p $(0 < p \leq 1)$, add s to the seed set $\mathcal{S}$. In Example Network, each of node 12's friends, nodes 1, 2, 8 and 9 (in green), are added probabilistically (with probability p) to the seed set. *Implementation:* For each $s \in \mathcal{N}_r$, draw from an independent uniformly distributed random variable $z_s \sim U[0, 1]$. If $z_s < p$, add s to the seed set $\mathcal{S}$.

Note: With Random and Local strategies, we will obtain exactly k nodes in the seed set $\mathcal{S}$. With the global strategy we might obtain more than k nodes in the seed set. In such a case, we select k nodes at random from the seed set $\mathcal{S}$ without replacement.



Example Network

The *global strategy* would ask individual nodes to provide a *proportion* of their friends, and such a proportion can be set to be small (say 1%) based on how many total seeds are required for interventions, and also to balance privacy concerns. The Global Strategy gives each friend of each random person an equal chance of being selected.[¶]

We illustrate how our approach is able to obtain the *relevant network structure* in a straightforward manner. Specifically, we query nodes to select from the *relevant network*. For instance, in application (a) where the focus was on physical contagion, the relevant network is the in person contact network. The query would then be phrased as "among the people you have interacted with in person, choose one at random." The idea of such queries to obtain the relevant network is general, and conditions can be added to the query (e.g. specifying a time period), depending on the desired intervention. Similar conditions can be used for applications (b) and (c). We can thus view such a query as providing the network that is relevant to the specific application.

Observe that with the local strategy, the number of seed nodes is fixed, whereas it is probabilistic under the global strategy. For the local strategy, by construction, the expected degree of the obtained seed is equal to the local mean. For the global strategy, we prove that the expected degree of chosen nodes is equal to the global mean (Theorem S5).

Though the local and global strategies appear to be similar in the sense that we are choosing friends of randomly chosen individuals, the crucial distinction lies in whether we are choosing *one* random friend (local) or a *proportion* of friends (global) of randomly chosen individuals, where the proportion is fixed in advance. This seemingly small distinction results in significant differences in both their mathematical properties, as well as empirically different quantities in real and generated networks. Table 1 details the algorithms to obtain k seeds in a network of size $N \gg k$. This impacts their relative effectiveness as examined below.

Effectiveness of Strategies: Leverage

To evaluate how much of an improvement over the random strategy is possible, and how this varies across a variety of generated and real networks, we examine the relative effectiveness of strategies, with the random strategy as the baseline and characterize leverage as the improvement in expected degree. Leverage for strategy s on network $\mathcal{G}$ is defined as $\lambda_s(\mathcal{G}) = \frac{\mu_s(\mathcal{G})}{\mu_D(\mathcal{G})}$ for $s \in \{R, L, G\}$ (since the random strategy obtains the mean degree in expectation, the leverage for R is $\lambda_R(\mathcal{G}) = 1$ and it serves as a baseline). We examine the leverage of both generated and real networks.

Generated Networks:. The generated networks using Random, Scale Free and Small World models are detailed in the Supplement §S.E. We examine 3 generative mechanisms: (a) Erdos-Renyi (ER), (b) Scale Free (SF) and (c) Small World (SW) networks (Fig. S5) (29–31). We find that for ER networks, at very low density (edge probability), the leverage is very low because most edges connect nodes that have a degree of 1. As density increases, we obtain more variation in degrees, and local leverage increases. However, beyond an edge probability of $p = 0.05$, leverage decreases as the density of the network decreases. Local leverage thus forms a non-monotonic pattern with ER networks. For SF networks, rather than density or edge probability, we initially examine leverage as the network becomes more centralized (as γ increases above 1, very high degree nodes have a lower probability of occurring). We find that as γ increases from 1 to 2, the leverage increases, but then decreases beyond 2. For WS networks, unlike in the ER and SF networks, leverage is monotonically decreasing with number of neighbors (or density), and is monotonically increasing with rewiring probability.

Real Networks:. The range of real networks is detailed in Supplement §S.C. First, observing the local strategy (Fig. 2A), we find that for all networks, as expected, the friendship paradox strategies are at least as good as the random strategy. Second, for networks like Twitter (OS4) or Internet Topology (C1), the leverage can be on the order of 100, implying that obtaining a connection of a random node will provide a 100-fold increase in expected degree. Third, we observe that both local and global leverage (Figs. 2A and 2B) are higher for nodes when average degree is intermediate, i.e. not too low or high. Some networks like the CA Roads network (I3) have very little degree variation and local and global strategies are relatively less effective. Finally, we examine when local and global strategies make a relative difference (Fig. 2B). We find that the highest ratio of local to global mean is for Twitter network (OS4), whereas the lowest ratio (indicating that global strategy has a higher expected mean degree) is shown by Flickr (OS2), both of which belong to the same category of online social networks. Citation networks tend to have higher global mean, whereas for Infrastructure networks, both strategies seem to work just as well.

Application: Controlling Contagion in Networks

We next illustrate the approach of using the friendship paradox strategies to obtain seeds for intervention, specifically vaccination in the face of simple contagion spreading through a network. Our goal is not for the application to directly inform immunization policy for a particular disease, but rather as a proof of concept. The virus propagation model here is simple and reduced to essential components. To be more realistic, the model would be more general, e.g. richer spatial models incorporating heterogeneity, potentially continuous and discrete time, and models that have parameters that are calibrated to match epidemiological data (32, 33).

We focus on simple models of contagion that can be characterized by a single parameter termed the *epidemic threshold* to focus our analysis on the benefit provided by the local and global strategies. If the ratio of infection to that of recovery is lower than the epidemic threshold, then the epidemic is contained and will die out, whereas if the ratio is above the threshold, then it could turn into an epidemic. The epidemic threshold captures the idea that a contagion introduced into the network will die out if the reproductive number ($\mathcal{R}_0$) is below the epidemic threshold, and lead to an epidemic if above the threshold. Thus, a network with a higher epidemic threshold would be able to better withstand or control an infection.

We then examine how the epidemic threshold changes as a function of the proportion of nodes vaccinated (removed),

[¶] These local and global strategies also have connections with respondent driven sampling (RDS), in which respondents nominate random friends or alters, e.g. by giving them participation tickets (28). An additional advantage of using such an approach is that the privacy risks are reduced further. The fact that these RDS based approaches have been commonly used in earlier interventions indicates that our proposed strategies are practical and knowledge about implementing them in specific contexts is likely to already exist.

using each strategy (random, local and global). For a wide class of *virus propagation models* (VPM), the epidemic threshold is characterized as the inverse of the greatest (first) eigenvalue of the adjacency matrix A of the network, denoted as $\tau(E) = \dfrac{1}{\lambda_1(E)}$ (details in the Supplement §S.H). The range of VPMs, including SIR, SEIR, etc. include models that have been commonly used for modeling infectious diseases (34). Nodes are selected for immunization or treatment using each of the intervention strategies (random, local and global).

We use in person contact networks for modeling contagion, with data on 75 village social networks from India (35). The social networks are captured at two different levels of aggregation, at the level of individuals and of households. The advantage of this dataset is that villages are relatively geographically isolated and can therefore be treated as separate networks. Details of the network dataset are provided in the Supplement§S.C.

We find that the village networks can have either positive or negative inversity depending on how nodes and edges are defined and aggregated. Figure 3(a) illustrates the inversity values across the 75 villages separately for individual and household networks. When nodes as defined as individuals, we find that the networks have negative inversity, whereas if the nodes are defined as households, the inversity values of the resulting networks are mostly positive. Inversity values therefore depend strongly on how the network structure is aggregated. Considering interventions, the inversity values suggest that a household-based intervention might use the local strategy, whereas the individual-based intervention might use the global strategy. Note that the household-level ties are aggregated from the individual-level ties, implying that networks obtained from similar underlying relationships can result in dramatically different inversity characteristics, which can lead to different interventions.

Our first step is to identify how the epidemic threshold τ changes as we immunize nodes from the network. This dataset is especially useful in our analysis since the villages are relatively isolated, implying they can be evaluated separately. While immunizing (or removing) any node from the network is likely to increase the epidemic threshold, immunizing highly-connected nodes is likely to prove especially beneficial. We examine the effectiveness of the three strategies (Random, Local and Global) in identifying which nodes to immunize in the network. In Figure 3(b), we show how the epidemic thresholds vary across strategies and proportion of nodes immunized (1% - 75%). In both household and individual networks, we find that the friendship paradox strategies obtain higher thresholds than random, for the same proportion of nodes immunized. For instance, in the household networks, to achieve a threshold $\tau = 0.15$, the random strategy needs to have about 50% of nodes immunized, but the Local and Global strategies require less than half of that, at around 25%. For the household networks, we find that the local strategy is better than the global strategies especially at higher levels of removal. However, for individual networks, we find that the Global strategy obtains greater thresholds than Local. This broadly signifies that it is helpful to know which among the global or local strategies to use, and is determined by the sign of inversity.

Finally, we simulate an infection process and evaluate the epidemic characteristic of peak infectivity using an SIR virus propagation model (details in §S.H)(36), with parameters of the simulation detailed in Table S5. We examine peak infection since it is known to be an important characteristic of epidemics (37), directly impacting the load on the healthcare system. Define $I_{it} \in \{0, 1\}$ as an indicator of whether an individual i is infected at time t. We evaluate the epidemics on the proportion of the population infected at the peak of the epidemic $\left(\frac{1}{N}\max_t(\sum_i I_{it})\right)$, which is a useful measure in cases where hospital capacity is constrained. There has been much discussion about interventions to avoid precisely such a peak (38). A strategy with a density plot to the left of another is better in terms of reducing the severity of the epidemic. Thus, for household networks, the local strategy (in red) is better than the global, which in turn is better than the random strategy in reducing peak infection. For the individual networks, however, the global strategy is better than the local strategy.

Conclusion

We have shown fundamental mathematical properties that underlie the friendship paradox, which we find to be multifaceted. We define and characterize the properties of the local mean, global mean and Inversity that connects the means for any network. We show that with unknown networks, the local and global strategies based on these means have theoretical guarantees on obtaining better-connected individuals from the relevant network. With both generated random networks and real networks, our results show the value of using the friendship paradox strategies to obtain highly connected nodes. In the vast majority of networks, we obtain at least double the average degree, and some networks show increases of close to a hundred-fold increase in node degree. We expect the advantages of these strategies, including sensitivity to privacy, speed of implementation, and generality of application areas, to prove important in using these strategies for interventions in unknown network structures.

Materials and Methods

Our analysis combines theoretical results along with simulation and empirical analysis on generated and real-world networks, in order to characterize the fundamental properties of the friendship paradox and related constructs.

Theoretical Properties. The Theoretical results are contained in SI Appendix S.B. We prove the Theorems on the individual friendship paradox, the properties of local mean and global means and inversity by using the properties of Networks (Graphs), and identifying the conditions required for these relationships to hold.

Empirical Analysis. For the empirical analysis, there are two separate but related parts. First, for the generated networks, in SI Appendix S.B, we examine the most commonly used generative mechanisms, i.e. Random Graphs (Erdos-Renyi), Scale Free (Barabasi-Albert) and Small World (Watts-Strogatz) networks. Second, real-world network data for simulations and empirical analysis is contained in SI Appendix S.C. Finally, we conducted a study of virus propagation under immunization carried out using the local, global and random strategies. The model specification, simulation, parameterization and values are contained in SI Appendix S.H. Simulation and empirical analysis was performed in R software, using igraph and sna packages.

ACKNOWLEDGMENTS. We acknowledge helpful comments from Steven Strogatz, the audiences at 2013 and 2016 Sunbelt Social Network Conferences, and seminar audiences at Massachusetts Institute of Technology, Yale University, University of Texas at Austin, University of California, San Diego and Washington University at St. Louis. The authors each acknowledge financial support from their respective universities.

1. S Feld, Why your friends have more friends than you do. *Am. J. Sociol.* pp. 1464–1477 (1991).
2. E Zuckerman, J Jost, What makes you think you're so popular? self-evaluation maintenance and the subjective side of the "friendship paradox". *Soc. Psychol. Q.* pp. 207–223 (2001).
3. HH Jo, YH Eom, Generalized friendship paradox in networks with tunable degree-attribute correlation. *Phys. Rev. E* **90**, 022809 (2014).
4. YH Eom, HH Jo, Generalized friendship paradox in complex networks: The case of scientific collaboration. *Sci. reports* **4**, srep04603 (2014).
5. DJ Higham, Centrality-friendship paradoxes: when our friends are more important than us. *J. Complex Networks* **7**, 515–528 (2019).
6. NO Hodas, F Kooti, K Lerman, Friendship paradox redux: Your friends are more interesting than you. *ICWSM* **13**, 8–10 (2013).
7. GT Cantwell, A Kirkley, M Newman, The friendship paradox in real and model networks. *J. Complex Networks* **9**, cnab011 (2021).
8. R Czaja, Sampling with probability proportional to size. *Encycl. Biostat.* **7** (2005).
9. A Nigam, P Kumar, V Gupta, Some methods of inclusion probability proportional to size sampling. *J. Royal Stat. Soc. Ser. B: Stat. Methodol.* **46**, 564–571 (1984).
10. C Orsini, et al., Quantifying randomness in real networks. *Nat. communications* **6**, 1–10 (2015).
11. T Valente, Network interventions. *Science* **337**, 49–53 (2012).
12. D Centola, M Macy, Complex contagions and the weakness of long ties. *Am. journal Sociol.* **113**, 702–734 (2007).
13. D Centola, The spread of behavior in an online social network experiment. *Science* **329**, 1194–1197 (2010).
14. R Cohen, S Havlin, D Ben-Avraham, Efficient immunization strategies for computer networks and populations. *Phys. review letters* **91**, 247901 (2003).
15. DA Kim, et al., Social network targeting to maximise population behaviour change: a cluster randomised controlled trial. *The Lancet* **386**, 145–153 (2015).
16. M Alexander, L Forastiere, S Gupta, NA Christakis, Algorithms for seeding social networks can enhance the adoption of a public health intervention in urban india. *Proc. Natl. Acad. Sci.* **119**, e2120742119 (2022).
17. D Kempe, J Kleinberg, É Tardos, Maximizing the spread of influence through a social network in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining.* pp. 137–146 (2003).
18. Z Katona, PP Zubcsek, M Sarvary, Network effects and personal influences: The diffusion of an online social network. *J. marketing research* **48**, 425–443 (2011).
19. C Wilson, A Sala, KP Puttaswamy, BY Zhao, Beyond social graphs: User interactions in online social networks and their implications. *ACM Transactions on Web (TWEB)* **6**, 1–31 (2012).
20. MA Al-Garadi, et al., Analysis of online social network connections for identification of influential users: Survey and open research issues. *ACM Comput. Surv. (CSUR)* **51**, 1–37 (2018).
21. A Acquisti, L Brandimarte, G Loewenstein, Secrets and likes: The drive for privacy and the difficulty of achieving it in the digital age. *J. Consumer Psychol.* **30**, 736–758 (2020).
22. F Cerruto, S Cirillo, D Desiato, SM Gambardella, G Polese, Social network data analysis to highlight privacy threats in sharing data. *J. Big Data* **9**, 19 (2022).
23. A Praveena, S Smys, Anonymization in social networks: a survey on the issues of data privacy in social network sites. *J. Int. J. Of Eng. And Comput. Sci.* **5**, 15912–15918 (2016).
24. W Xie, K Karan, Consumers' privacy concern and privacy protection on social network sites in the era of big data: Empirical evidence from college students. *J. Interact. Advert.* **19**, 187–201 (2019).
25. S Torabi, K Beznosov, Privacy aspects of health related information sharing in online social networks in *2013 USENIX Workshop on Health Information Technologies (HealthTech 13).* (2013).
26. B Liu, et al., When machine learning meets privacy: A survey and outlook. *ACM Comput. Surv. (CSUR)* **54**, 1–36 (2021).
27. S Wachter, Normative challenges of identification in the internet of things: Privacy, profiling, discrimination, and the GDPR. *Comput. law & security review* **34**, 436–449 (2018).
28. SK Thompson, *Sampling.* (John Wiley & Sons) Vol. 755, (2012).
29. P Erdõs, A Rényi, On random graphs, i. *Publ. Math. (Debrecen)* **6**, 290–297 (1959).
30. AL Barabási, R Albert, Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
31. D Watts, S Strogatz, Collective dynamics of "small-world" networks. *nature* **393**, 440–442 (1998).
32. LJ Thomas, et al., Spatial heterogeneity can lead to substantial local variations in covid-19 timing and severity. *Proc. Natl. Acad. Sci.* **117**, 24180–24187 (2020).
33. LJ Thomas, et al., Geographical patterns of social cohesion drive disparities in early covid infection hazard. *Proc. Natl. Acad. Sci.* **119**, e2121675119 (2022).
34. BA Prakash, D Chakrabarti, M Faloutsos, N Valler, C Faloutsos, Got the flu (or mumps)? check the eigenvalue! *arXiv preprint arXiv:1004.0060* (2010).
35. A Banerjee, AG Chandrasekhar, E Duflo, MO Jackson, The diffusion of microfinance. *Science* **341**, 1236498 (2013).
36. J Tolles, T Luong, Modeling epidemics with compartmental models. *Jama* **323**, 2515–2516 (2020).
37. A Soria, et al., The high volume of patients admitted during the SARS-Cov-2 pandemic has an independent harmful impact on in-hospital mortality from covid-19. *PloS one* **16**, e0246170 (2021).
38. RD Booton, et al., Estimating the COVID-19 epidemic trajectory and hospital capacity requirements in South West England: a mathematical modelling framework. *BMJ Open* **11**, e041536 (2021).

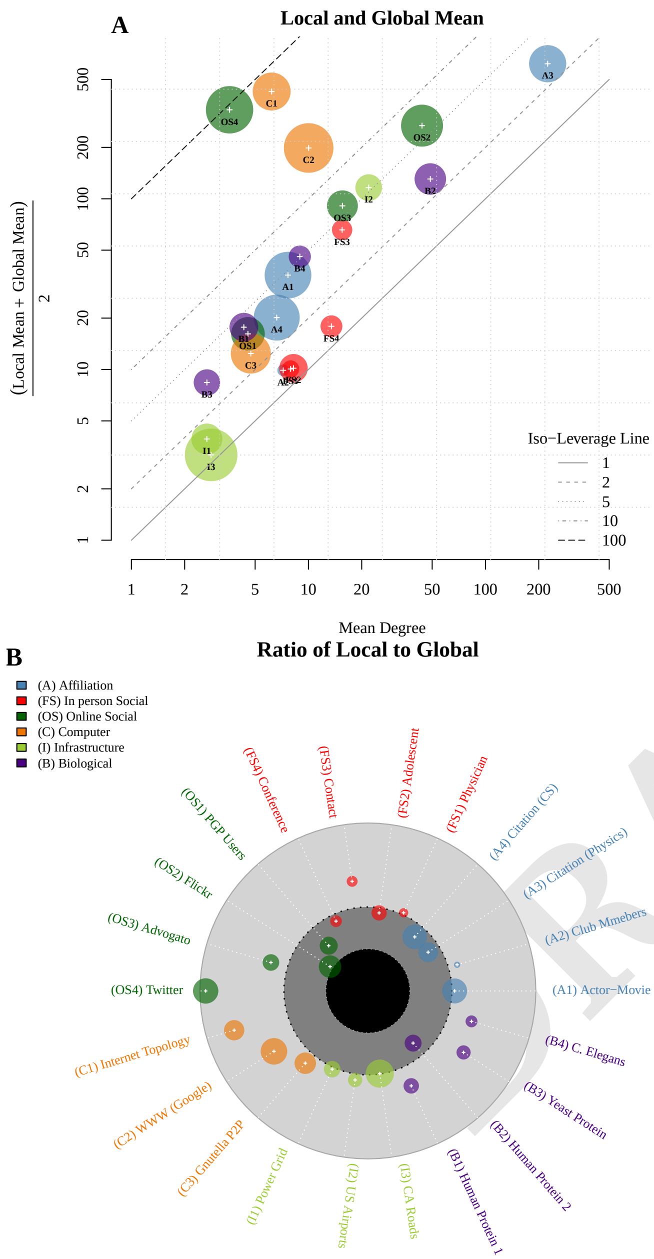


Fig. 2. Global and Local Leverage in Real Networks Local and Global Means across Networks (each circle is a network). Area of circles indicates size of networks (number of nodes) in log scale. Color of circle indicates network category. (A) The average of **Local and Global Mean** is higher than mean degree in all real networks, with the highest differences occurring in online social networks and computer networks. Most large networks also tend to show a higher leverage ratio. For in person or face to face networks, the pattern is more variable. The iso-leverage line indicates leverage levels of 1,2,5,10 and 100. We find that all networks have leverage greater than 1, a majority of networks have leverage greater than 5, and 2 networks have leverage close to 100. (B) **Comparison**: Ratio of Local to Global Mean. The ratio of local to global mean $\frac{\mu_L}{\mu_G}$ is represented as follows ($< \frac{1}{2}$ in black circle, $\frac{1}{2} < \frac{\mu_L}{\mu_G} < 1$ in dark gray circle and $1 < \frac{\mu_L}{\mu_G} < 2$ in light gray circle. For example, in the Twitter network, local mean is almost twice the global mean, whereas in the Flickr network, global mean is almost twice the local mean. Computer networks have higher values of the ratio, whereas Infrastructure networks have similar values of local and global means.

Fig. 3. Inversity and Epidemic Characteristics on Village Networks. Data for these networks (N=75) obtained from Indian villages is publicly available and detailed in (35), and includes both individual-level (Individual network) ties as well as connections between households (Household network). (a): Inversity: Frequency plot of inversity across village networks. (b) Epidemic Thresholds with Immunization: The epidemic threshold $\left(\tau = \frac{1}{\lambda_1}\right)$ is computed at different levels of immunization, and for different immunization strategies (Random, Local and Global). The strategies are used to select nodes, which upon immunization lose the capability to transmit contagion, and infect other nodes. (c) Epidemic Peak Infection: The proportion of the network nodes that are infected during the peak of the epidemic for immunization level (X) is represented as a density plot. Variation is obtained due to differences in outcomes across village as well as simulation variation. See Table S5 in the Supplement for simulation parameters.