On the friendship paradox and inversity: A network property with applications to privacy-sensitive network interventions

Vineet Kumar[a,1] (ID), David Krackhardt[b] (ID), and Scott Feld[c] (ID)

We provide the mathematical and empirical foundations of the friendship paradox in networks, often stated as "Your friends have more friends than you." We prove a set of network properties on friends of friends and characterize the concepts of ego-based and alter-based means. We propose a network property called inversity that quantifies the imbalance in degrees across edges and prove that the sign of inversity determines the ordering between ego-based or alter-based means for any network, with implications for interventions. Network intervention problems like immunization benefit from using highly connected nodes. We characterize two intervention strategies based on the friendship paradox to obtain such nodes, with the alter-based and ego-based strategy. Both strategies provide provably guaranteed improvements for any network structure with variation in node degrees. 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 works better based on network topology and show how network aggregation can alter inversity. We illustrate how the 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, are flexible for time-sensitive action, and only require selected nodes to nominate network neighbors.

networks | contagion | seeding

We examine the underlying mathematical and empirical foundations of the friendship paradox and define a 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 distinct ways of understanding this statement, which lead to different network properties that we term as the ego-based and alter-based 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 so across a wide class of generated and real-world networks. We identify a network property, inversity, that connects the two means, and for any network, the sign of inversity determines whether the ego-based mean or alter-based mean is higher. Since these mathematical properties apply to any network, not just those based on friendship, we use neighbor in place of friend henceforth.

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 ego-based and alter-based strategies for obtaining highly connected nodes, of which the alter-based strategy has not been used in network interventions before. The inversity of the network indicates which strategy (ego-based or alter-based) is better for obtaining highly connected nodes. We show in empirical real-world networks, and in generated networks, that the strategy used can make a meaningful difference. We illustrate, using a simplified application with real networks, how using inversity to choose the ego-based or alter-based strategy to identify inoculation candidates considerably reduces 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 neighbors are more popular than them (1, 2).* The intuition for why

Significance

Networks across many different settings—including social, economic, and natural—are powerful tools for interventions due to the cascading impact of one individual node on others. All networks with degree variation exhibit the friendship paradox phenomenon. We demonstrate its multifaceted nature, and provide its foundations mathematically and empirically. We identify a network property—inversity—and propose network intervention strategies based on the friendship paradox. Inversity uniquely determines the best-performing strategy. These strategies provide a privacy-sensitive approach to obtaining highly connected individuals without knowing the network, and are guaranteed to obtain a greater than average degree for almost any network. Finally, we characterize the value of these strategies theoretically and with real-world networks.

---

* The phenomenon has also been generalized to the idea that individual attributes and degree are correlated (3), e.g., an individual's coauthors are more likely to be cited (4), or that neighbors are more important (5), or more socially active (6). Of specific relevance to this research is the mathematical generalization to distributions examined in ref. 7.

the friendship paradox helps obtain well-connected nodes is this: Since highly connected nodes (hubs) are connected to many other nodes, obtaining a random friend (neighbor) of a random node is likely to result in selecting hubs with greater likelihood, compared to the case of randomly selecting nodes.

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

Ego-Based and Alter-Based Means. We formally characterize the two distinct but related network properties deriving from the friendship paradox relating to the "average number of neighbors of neighbors." We denote an undirected network (see *SI Appendix*, Table S1 for full notation) as a graph $\mathcal{G} = (V, E)$ with V the set of vertices or 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 $\mathcal{N}(i)$ refers to the set of i's neighbors. We specify the ego-based mean as the average number of neighbors of neighbors across the nodes in the network, consistent with (8):

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

The alter-based mean is defined as the ratio of the total number of neighbors of neighbors to the total number of neighbors in the network, consistent with (1):

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

The above means arise from conceptualizing the average degree across neighbors differently.[†] Both means above are consistent with the notion of "average number of neighbors of neighbors," although they are distinct network properties (see *SI Appendix*, Fig. S1 for an example and detailed explanation). The alter-based mean was theoretically investigated earlier and found to be greater than (or equal to) the average degree and is independent of the network topology, given node degrees (*SI Appendix*, Theorem S2). Equality holds only when the network is regular, with all nodes the same degree within and across components.[‡] The ego-based mean is shown here to be

[†] We note that a node is also its neighbor's neighbor under both ego-based and alter-based properties.

[‡] 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 a class size that is larger than the average size of classes. The underlying selection process here is commonly termed probability proportional to size (PPS) (9, 10). In the case of the friendship paradox, we find that the alter-based mean has a direct mathematical connection to PPS (neighbors have disproportionately more neighbors). However, the ego-based mean operates through a different mechanism.

greater than (or equal to) the mean degree (*SI Appendix*, Theorem S3).[§] However, the contrast is that the ego-based mean has distinct properties that depend on network topology (i.e., who is connected to whom). Equality for the ego-based mean only holds when each component is regular, with no degree variation within components.

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

Inversity: Connecting Ego-Based and Alter-Based Means

We identify and define a network property, inversity, that determines when the ego-based mean is greater than the alter-based mean. This property captures all local network information related to the ego-based mean. We prove that the sign of inversity determines whether the ego-based mean or the alter-based mean is higher for any given network.

Inversity is a correlation-based metric that relates the alter-based and ego-based means for any network, and 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.

$$\rho = \text{Cor}\left(D^{\mathbf{O}}, D^{\mathbf{ID}} \right) \quad [3]$$

We show that the ego-based and alter-based means are connected by *inversity* and the degree distribution ($\kappa_m = \sum_{i \in V} D_i^m$) as follows:

$$\mu_E = \mu_A + \rho \, \Psi(\kappa_{-1}, \kappa_1, \kappa_2, \kappa_3) \quad [4]$$

where Ψ is a positive function of moments of the degree distribution (*SI Appendix*, Theorem S4).

Inversity is a measure of *imbalance* between the degrees of the nodes connected by an edge. This imbalance for edge (i, j) is characterized by the ratio of degrees in both directions, i.e. $\left(\frac{D_i}{D_j} \right)$ and $\left(\frac{D_j}{D_i} \right)$. That is, more imbalanced edges tend to have nodes with a high degree on one end and a low degree on the other end. This imbalance plays a significant role in both inversity and the ego-based mean. Consider how the ego-based mean is obtained: For each node, we take the mean of degrees across the node's neighbors. The expected degree of neighbors of nodes (or ego-based mean) is then $\frac{1}{N} \sum_i \sum_{j \in \mathcal{N}(i)} \left(\frac{D_i}{D_j} \right)$. This imbalance

[§] We term these means ego-based or alter-based since they reflect the experience of nodes in their roles as ego or alter respectively. See *SI Appendix*, §S.A.

measure based on ratios explains why inversity is highly sensitive to stars, whereas assortativity is more sensitive to the presence of clique-like structures.

When connections (edges) are mostly between nodes of similar degree, then inversity ρ is likely to be more negative. In such a case, the alter-based mean is greater than the ego-based mean. In contrast, when connections are more likely to be between nodes of dissimilar degree, then inversity ρ is positive, and the ego-based strategy is likely to obtain higher degree nodes. If inversity is known, we do not need the entire degree distribution to obtain the ego-based mean. Rather, *four* moments of the degree distribution are sufficient for that purpose. Inversity captures the local information on imbalances in the degrees 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 ego-based or alter-based mean is greater for a network; specifically, $\rho < 0$ indicates the alter-based mean is higher than the ego-based 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 (11).

How Inversity Depends On Network Topology. Inversity is the only term that depends on the network topology in the relationship between ego-based and alter-based means. We examine how inversity changes as we change the topology, while *simultaneously* preserving the degree distribution. In Fig. 1, we start with a network with the minimum inversity $\rho = -1$, and then use the rewiring theorem (*SI Appendix,* Theorem S6) to examine how inversity increases while the degree distribution remains fixed. Consequently, the mean degree, variance of degree, minimum and maximum degree, as well as the alter-based mean all remain fixed and are identical across each of the networks (A–I). Specifically, each network has the following properties in common: $N = 25$ nodes, $\mu_D = 3.5$, $\sigma_D^2 = 3.4$, $D_{\min} = 1$, and $D_{\max} = 8$. Observe that the alter-based mean, $\mu_A = 6.7 > \mu_D = 3.5$, is also the same for each of the networks.

The motivation for keeping degree distribution fixed across the networks is to examine how local network topology impacts only the ego-based mean, and not the alter-based mean. 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 degrees, which is essential for inversity to be greater than its minimum value.

We use the rewiring theorem (*SI Appendix,* 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 and B) and in each further step. We observe that the nodes in each dyad break up their edge (which connects nodes of identical degrees), 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 panels (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 the ego-based mean can be highly sensitive to
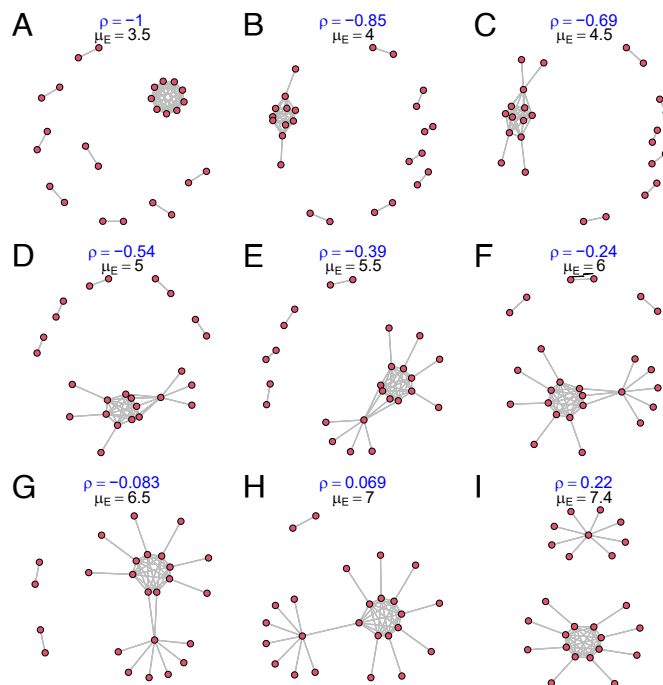

Fig. 1. How inversity changes with rewiring. The network is changed by rewiring, starting with the *Top Left* ($\rho = -1$), to increase inversity ρ as we traverse from panel (A–I). Observe that the number of nodes, $N = 25$, and the number of edges $|E| = 44$, as well as the degree distribution for 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 ego-based mean is $\mu_E = \mu_D = 3.5$ for network (A), but increases along with inversity in panels (B–I), reaching $\mu_E = 7.4$ for network (I). The alter-based mean, $\mu_A = 6.7$, remains constant across all the networks.

network topology. 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 with different levels of inversity and ego-based mean for a fixed alter-based mean. These inversity values range from negative to positive. Finally, the degree distribution can constrain the range of inversity. We next examine the implications of these findings for network interventions.

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 a few ties that largely go to the relatively few hubs with large numbers of ties (e.g. social media network of a celebrity). In contrast, we expect that networks of clusters of various sizes will have low inversity. In such clusters, members of the large clusters are tied to one another, obtaining high degree nodes within the cluster. Similarly, nodes in small clusters tend to have few ties, mostly with one another, reflecting 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 *SI Appendix,* §S.F. For details on inversity values in real networks, see *SI Appendix,* §S.C.

Inversity and Assortativity. Inversity $\rho = \text{cor}(D_i, \frac{1}{D_j})$ is related to, but distinct from, the commonly used measure of assortativity, defined as $\rho_a = \text{cor}(D_i, D_j)$. We observe that the formulation of both network properties appears similar, and both have values ranging from -1 to $+1$. We might therefore expect inversity to be a reverse of assortativity, or more specifically $\rho_a \approx -\rho$.

Can assortativity serve as a proxy for inversity? [¶] We show that inversity and assortativity can both be the same sign, i.e. both positive (or both negative). We demonstrate that this same-sign property can hold across a range of networks, with the effects detailed in *SI Appendix* §S.G. This leads to cases where reliance on assortativity as a proxy for inversity would lead to incorrect decisions about optimal intervention strategies. Next, these decisions could also require knowing not just the sign of inversity, but the difference in magnitudes of expected degree across the intervention strategies. Approximating $\rho \approx -\rho_a$ to obtain expected degree differences would further magnify these errors. Broadly, we document how using assortativity as a proxy metric for inversity is not conceptually appropriate, and not required since inversity is equally easy to compute. However, it would be useful to further examine this distinction for real-world networks.

No monotonic ordering. We observe that there are many pairs of networks for which we do not obtain appropriate monotonic ordering across inversity and assortativity. Specifically, there are pairs of networks $\mathcal{G}_1$ and $\mathcal{G}_2$, such that we have $\rho(\mathcal{G}_2) > \rho(\mathcal{G}_1)$ and $\rho_a(\mathcal{G}_2) > \rho_a(\mathcal{G}_1)$, so *both* assortativity and inversity are higher for one of the networks. Such an ordering should not be possible to obtain if assortativity and inversity are a reverse ordering of each other. The implication is that there is not a clear one-to-one mapping between assortativity and inversity in networks.

Network Interventions

Consider the following problems: (a) (Reducing) A new infectious disease is spreading through a large population. We want to minimize the number of infected individuals by inoculating the population using a new vaccine. However, we only have a limited number of doses to administer. (b) (Accelerating) We have a highly effective medical device with limited samples that we would like to provide to select medical professionals, who can then share the information through word-of-mouth. (c) (Observing) We would like to identify a viral contagion as quickly as possible by choosing individuals as observation stations (or for contact tracing). Although seemingly disparate, these problems of how to reduce, accelerate, or observe dynamic contagion represent a class of network interventions in which we benefit from identifying more central or highly connected individuals in the network (12).[#]

We show how seeding interventions using the friendship paradox, using the ego-based and alter-based strategies, impact network interventions by helping to obtain highly connected seed nodes in a privacy-sensitive manner from the relevant network. Note that while the ego-based strategy has been commonly suggested and used in practice (15–17), its theoretical and empirical properties have not been examined and characterized for general networks. The alter-based strategy is novel, 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 (18–20). In such cases, privacy concerns are increasingly relevant, making it challenging to obtain network data (21–25). Users are also

concerned that their data may be used in algorithms (26) and even result in discrimination against them (27). There have been methods proposed to use only local knowledge and leverage the community structure to identify influential nodes to use in network interventions for disease control, so that full knowledge of the network topology is not required. Such approaches can be used to find bridging nodes that span communities (28). Other have suggested nodes that are influential within communities when considering the community structure (29), which arise from mechanisms like propinquity (30). However, obtaining theoretical guarantees without full knowledge of the network structure is challenging. Our approach, in contrast, is focused on obtaining higher degree nodes, with provable guarantees for general network topologies.

Relevant Network Topology. When obtaining network topology, the challenge in many cases is that we do not have access to the *relevant* network. For instance, 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 emphasize the importance of being able to leverage the structure of the relevant network, while being sensitive to privacy concerns.

The friendship paradox based strategies have several advantages for implementation. First, despite being informationally light, the strategies here provide provable advantages for virtually any network topology, in contrast to strategies that do not provide such guarantees for general networks. The network structure may also be expensive to collect, may 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 fully 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.

Implementing the Intervention Seeding Strategies

The above formulations of the ego-based and alter-based means suggest distinct strategies for choosing seeds for interventions, or intervention strategies. We illustrate *random*, *ego-based*, and *alter-based* strategies to choose a "seed" node in the network beginning with an initial randomly chosen node (Table 1). For example, the *ego-based strategy* would query randomly selected individual nodes with the query, "could you suggest the name of a randomly chosen neighbor?"

The *alter-based strategy* would ask individual nodes to choose each of their neighbors with a probability that is fixed across nodes. The probability can be set to be small (say $P = 0.05$) based on how many total seeds are required for interventions, and also to balance privacy concerns. The alter-based strategy gives each neighbor of each random person an equal chance of being selected, and we prove that the expected degree of chosen nodes


[¶]For details of each of these arguments and examples, please see *SI Appendix*, §S.G.

[#]We focus on the class of "simple contagion" problems, which require only one exposure, rather than "complex contagion" problems, which require multiple exposures (13, 14).

Table 1. Implementation of seeding strategies

Step	Details
0	Fix $P \in (0, 1]$ (only used for alter-based strategy in Step 2G).
	Repeat Steps 1 to 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. If r is an isolate, repeat this step.
	In example network, Nodes 10, 18, and 12 (in black) are drawn for (R), (E) and (A) strategies respectively.
2	Depending on the strategy Random (R), Ego-based (E), or Alter-based (A), do the following:
	2R (Random): Add r to the seed set $\mathcal{S}$.
	In example network, add node 10 to the seed set.
	2L (Ego-based): Obtain a node s chosen with uniform probability from r's neighbors, i.e., $s \in \mathcal{N}_r$. Add the neighbor s to the seed set $\mathcal{S}$.
	In example network, one of node 18's neighbors, node 14 (in red), is chosen at random. Add node 14 to the seed set.
	2G (Alter-based): For each of r's neighbors, $s \in \mathcal{N}_r$: with probability P ($0 < P \le 1$), add s to the seed set $\mathcal{S}$.
	In example network, each of node 12's neighbors, 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 Ego-based strategies, we will obtain exactly k nodes in the seed set $\mathcal{S}$. With the Alter-based 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
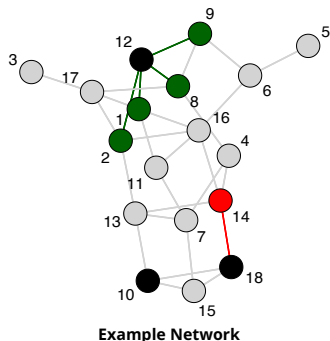
is equal to the alter-based mean (*SI Appendix*, Theorem S5).$^{\|}$ The crucial distinction between the ego-based and alter-based strategies lies in whether we are choosing *one* random neighbor (ego-based) or a fixed probability for *each* neighbor (alter-based) of randomly chosen individuals. Depending on the application, these strategies could be different in ease of implementation, with ego-based strategy appearing to be simpler. Table 1 details the algorithms to obtain k seeds in a network of size $N \gg k$.

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

$^{\|}$These ego-based and alter-based strategies also have connections with respondent driven sampling (RDS), in which respondents nominate random neighbors or alters, e.g., by giving them participation tickets (31). 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.

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 the above as a query that provides a network that is relevant to the specific application. Next, we evaluate the relative effectiveness of these intervention strategies.

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 $s \in \{R, E, A\}$ (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 were obtained using three generative mechanisms (32–34): (a) Random or Erdos–Renyi (ER), (b) Scale Free (SF), and (c) Small World (SW) models, as detailed in *SI Appendix*, §S.E.

The results are detailed in *SI Appendix*, Fig. S5. 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 ego-based leverage increases. However, beyond an edge probability of $P = 0.05$, leverage decreases as the density of the network increases. Ego-based leverage thus forms a nonmonotonic 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 SW 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 *SI Appendix*, §S.C. First, observing the ego-based strategy (Fig. 2*A*), 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 highly substantial—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 ego-based and alter-based leverage (Fig. 2 *A* and *B*) 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 ego-based and alter-based strategies are relatively less effective. Finally, we examine the conditions under which ego-based and alter-based strategies have a relative difference (Fig. 2*B*). We find that the highest ratio of ego-based to alter-based mean is for the Twitter network (OS4), whereas the lowest ratio (indicating that the alter-based strategy has a higher expected mean degree) is shown by Flickr (OS2), both of which belong to the same category of online
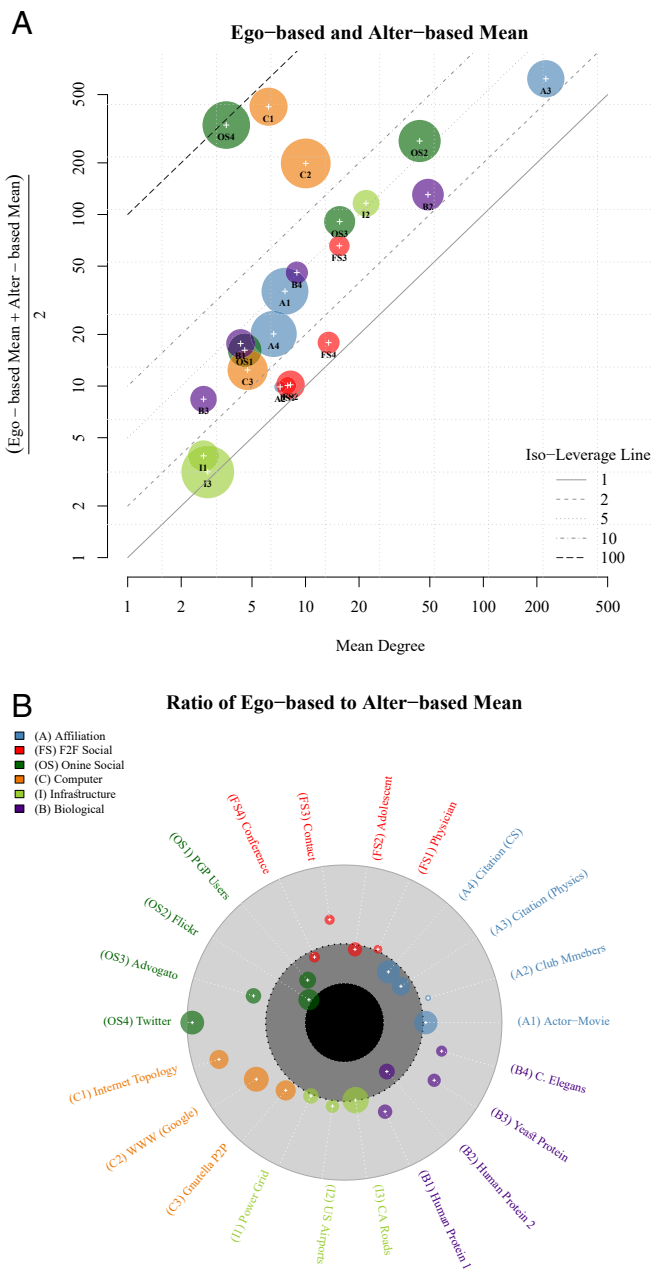
A

Ego-based and Alter-based Mean

B

Ratio of Ego-based to Alter-based Mean

Fig. 2. Alter-based and ego-based leverage in real networks ego-based and alter-based means across networks (each circle is a network). Area of circles indicates size of networks (number of nodes) in log scale. Color of the circle indicates network category. (*A*) The average of ego-based and alter-based 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 isoleverage 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 ego-based to alter-based mean. The ratio of ego-based to alter-based mean $\frac{\mu_E}{\mu_A}$ is represented as follows ($< \frac{1}{2}$ in black circle, $\frac{1}{2} < \frac{\mu_E}{\mu_A} < 1$ in dark gray circle and $1 < \frac{\mu_E}{\mu_A} < 2$ in light gray circle. For example, in the Twitter network, ego-based mean is almost twice the alter-based mean, whereas in the Flickr network, alter-based mean is almost twice the ego-based mean. Computer networks have higher values of the ratio, whereas Infrastructure networks have similar values of ego-based and alter-based means.

social networks. Citation networks tend to have a higher alter-based 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 to serve 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 having parameters calibrated to match epidemiological data (35, 36).

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 ego-based and alter-based strategies. The epidemic threshold captures the idea that a contagion in the network will die out if the reproductive number ($\mathcal{R}_0$) is below the epidemic threshold, and will lead to an epidemic if $\mathcal{R}_0$ is 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), using each strategy (random, ego-based, and alter-based).

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 of the network, denoted as $\tau(E) = \frac{1}{\lambda_1(E)}$ (details in *SI Appendix*, §S.H). The above formulation applies to a range of VPMs, including SIR, SEIR, etc., which include models commonly used for infectious diseases (37).
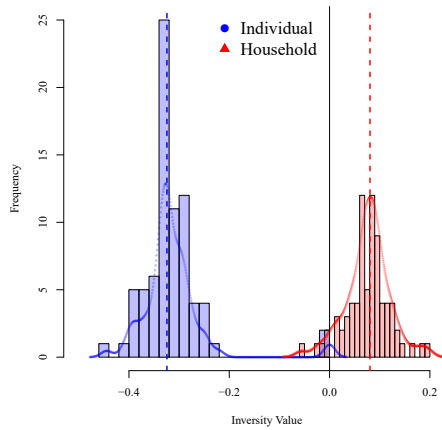
We use in-person contact networks for modeling contagion, with data on 75 village social networks from India (38). The social networks are captured at two different levels of aggregation, at the level of individuals and of households. The advantage of this dataset is that villages are relatively geographically isolated and can therefore be treated as separate networks. Details of the network dataset are provided in *SI Appendix*, §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. Fig. 3*A* illustrates the inversity values across the 75 villages separately for individual and household networks. When nodes are 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. Since household-level ties are aggregated from the individual-level ties, we find that networks obtained from similar underlying relationships can result in dramatically different inversity characteristics, which can lead to different interventions. Considering interventions, the inversity values suggest that a household-based intervention might use the ego-based strategy, whereas the individual-based intervention might use the alter-based strategy.
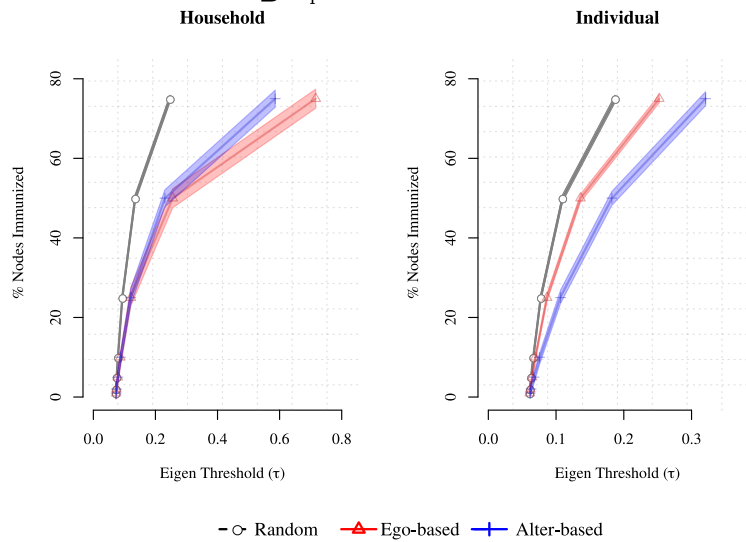
We next evaluate how the epidemic threshold τ changes as we immunize nodes from the network for each of the strategies (random, ego-based, and alter-based). While immunizing (or removing) any node from the network is likely to increase the epidemic threshold, immunizing highly connected nodes is likely to prove especially beneficial. In Fig. 3*B*, we show how the epidemic thresholds vary across strategies and proportion of nodes immunized (1% to 75%). In both household and individual networks, we find that the friendship paradox strategies obtain higher epidemic thresholds than the random strategy,
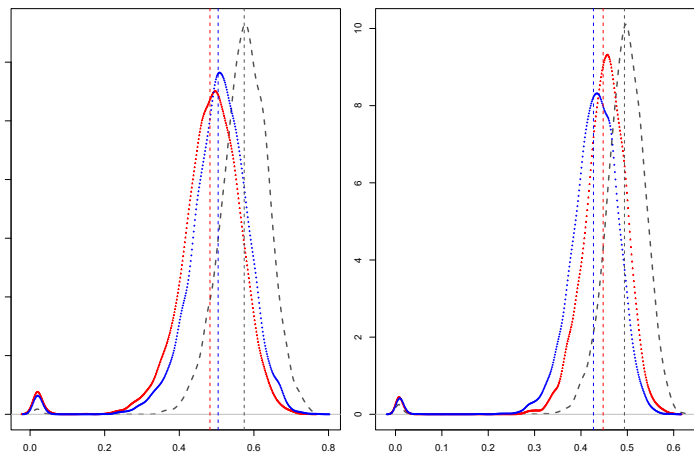
Fig. 3. Inversity and epidemic characteristics on village networks. Data for these networks ($N = 75$) obtained from Indian villages are publicly available and detailed in ref. 38. The data include both individual-level (individual network) ties as well as connections between households (household network). (*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, ego-based, and alter-based). 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 is represented as a density plot. Variation is obtained due to differences in outcomes across villages as well as simulation variation. See *SI Appendix*, Table S5 for simulation parameters. Note that in panels (*B* and *C*), the red curves correspond to the ego-based strategy, whereas the blue curves correspond to the alter-based strategy.

for the same proportion of nodes immunized. For instance, in the household networks, to achieve a epidemic threshold $\tau = 0.15$, the random strategy needs to have about 50% of nodes immunized, but the ego-based and alter-based strategies require less than half of that, at around 25%. For the household networks, we find that the ego-based strategy is better than the alter-based strategy, especially at higher levels of removal. However, for individual networks, we find that the alter-based strategy obtains greater thresholds than the ego-based strategy. This broadly signifies that it is helpful to know which among the alter-based or ego-based strategies to use, as determined by the sign of inversity.

Finally, we simulate an infection process and evaluate the epidemic characteristic of peak infection using an SIR virus propagation model (details in *SI Appendix*, §S.H) (39), with parameters of the simulation detailed in *SI Appendix*, Table S5. We examine peak infection since it is known to be an important characteristic of epidemics (40), directly impacting the load on the healthcare system. We denote $I_{it} \in \{0, 1\}$ as an indicator of whether an individual i is infected at time t. We evaluate 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 healthcare capacity is constrained. There has been much discussion about the value of interventions to avoid and minimize such a peak (41). 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 ego-based strategy (in red) is better than alter-based, which in turn is better than the random strategy in reducing peak infection. For individual networks, however, the alter-based strategy is better than the ego-based strategy. Overall, we find that friendship paradox strategies (ego-based and alter-based) clearly improve upon the random 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 ego-based mean, alter-based mean, and inversity to connect the means for any network. We show that for unknown networks, the ego-based and alter-based 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 substantial value of using the friendship paradox strategies to obtain highly connected nodes. In the vast majority of networks, these strategies 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 concerns, speed of implementation, and generality of application areas, to be important factors in using them 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 ego-based mean and alter-based means, and inversity by using the properties of networks and identifying the conditions under which these relationships hold.

Empirical Analysis. For the empirical analysis, there are two separate but related parts. First, for the generated networks, in *SI Appendix*, §S.E, 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, we use real-world network data for simulations and include the empirical analysis in *SI Appendix*, §S.C. Finally, we conducted a study of virus propagation under immunization carried out using the ego-based, alter-based, 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.

1. S. Feld, Why your friends have more friends than you do. *Am. J. Sociol.*, 1464–1477 (1991).
2. E. W. Zuckerman, J. T. Jost, What makes you think you're so popular? Self-evaluation maintenance and the subjective side of the "friendship paradox" *Soc. Psychol. Q.* **64**, 207–223 (2001).
3. H. H. Jo, Y. H. Eom, Generalized friendship paradox in networks with tunable degree-attribute correlation. *Phys. Rev. E* **90**, 022809 (2014).
4. Y. H. Eom, H. H. Jo, Generalized friendship paradox in complex networks: The case of scientific collaboration. *Sci. Rep.* **4**, srep04603 (2014).
5. D. J. Higham, Centrality-friendship paradoxes: When our friends are more important than us. *J. Complex Netw.* **7**, 515–528 (2019).
6. N. O. Hodas, F. Kooti, K. Lerman, Friendship paradox redux: Your friends are more interesting than you. *ICWSM* **13**, 8–10 (2013).
7. G. T. Cantwell, A. Kirkley, M. Newman, The friendship paradox in real and model networks. *J. Complex Netw.* **9**, cnab011 (2021).
8. D. Krackhardt, Structural leverage in marketing. Networks in marketing (1996), pp. 50–59.
9. R. Czaja, "Sampling with probability proportional to size" in *Encyclopedia of Biostatistics* (2005), vol. 7.
10. A. Nigam, P. Kumar, V. Gupta, Some methods of inclusion probability proportional to size sampling. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **46**, 564–571 (1984).
11. C. Orsini *et al.*, Quantifying randomness in real networks. *Nat. Commun.* **6**, 1–10 (2015).
12. T. Valente, Network interventions. *Science* **337**, 49–53 (2012).
13. D. Centola, M. Macy, Complex contagions and the weakness of long ties. *Am. J. Sociol.* **113**, 702–734 (2007).
14. D. Centola, The spread of behavior in an online social network experiment. *Science* **329**, 1194–1197 (2010).
15. R. Cohen, S. Havlin, D. Ben-Avraham, Efficient immunization strategies for computer networks and populations. *Phys. Rev. Lett.* **91**, 247901 (2003).
16. D. A. Kim *et al.*, Social network targeting to maximise population behaviour change: A cluster randomised controlled trial. *Lancet* **386**, 145–153 (2015).
17. M. Alexander, L. Forastiere, S. Gupta, N. A. Christakis, Algorithms for seeding social networks can enhance the adoption of a public health intervention in urban India. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2120742119 (2022).
18. 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* (2003), pp. 137–146.
19. Z. Katona, P. P. Zubcsek, M. Sarvary, Network effects and personal influences: The diffusion of an online social network. *J. Mark. Res.* **48**, 425–443 (2011).
20. M. A. Al-Garadi *et al.*, Analysis of online social network connections for identification of influential users. *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. Consum. Psychol.* **30**, 736–758 (2020).
22. F. Cerruto, S. Cirillo, D. Desiato, S. M. 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. Eng. 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* (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 Sec. Rev.* **34**, 436–449 (2018).

28. M. Salathé, J. H. Jones, Dynamics and control of diseases in networks with community structure. *PLoS Comput. Biol.* **6**, e1000736 (2010).

29. N. Gupta, A. Singh, H. Cherifi, Centrality measures for networks with community structure. *Phys. A Stat. Mech. Appl.* **452**, 46–59 (2016).

30. L. K. Gallos, S. Havlin, H. E. Stanley, N. H. Fefferman, Propinquity drives the emergence of network structure and density. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 20360–20365 (2019).

31. S. K. Thompson, *Sampling* (John Wiley & Sons, 2012), vol. 755.

32. P. Erdös, A. Rényi, On random graphs I. *Publ. Math. Debrecen* **6**, 290–297 (1959).

33. A. L. Barabási, R. Albert, Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).

34. D. Watts, S. Strogatz, Collective dynamics of "small-world" networks. *Nature* **393**, 440–442 (1998).

35. L. J. Thomas *et al.*, Spatial heterogeneity can lead to substantial local variations in COVID-19 timing and severity. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 24180–24187 (2020).

36. L. J. Thomas *et al.*, Geographical patterns of social cohesion drive disparities in early Covid infection hazard. *Proc. Natl. Acad. Sci. U.S.A.* **119**, e2121675119 (2022).

37. B. A. Prakash, D. Chakrabarti, M. Faloutsos, N. Valler, C. Faloutsos, Got the flu (or mumps)? check the eigenvalue! arXiv [Preprint] (2010). http://arxiv.org/abs/1004.0060.

38. A. Banerjee, A. G. Chandrasekhar, E. Duflo, M. O. Jackson, The diffusion of microfinance. *Science* **341**, 1236498 (2013).

39. J. Tolles, T. Luong, Modeling epidemics with compartmental models. *JAMA* **323**, 2515–2516 (2020).

40. 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).

41. R. D. 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).

42. J. Kunegis, "Konect: the koblenz network collection" in *Proceedings of the 22nd International Conference On World Wide Web* (2013), pp. 1343–1350.

PNAS

Supporting Information for

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

Vineet Kumar, David Krackhardt and Scott Feld

Corresponding Author: Vineet Kumar.
E-mail: vineet.kumar@yale.edu

This PDF file includes:

Figs. S1 to S8
Tables S1 to S5
SI References

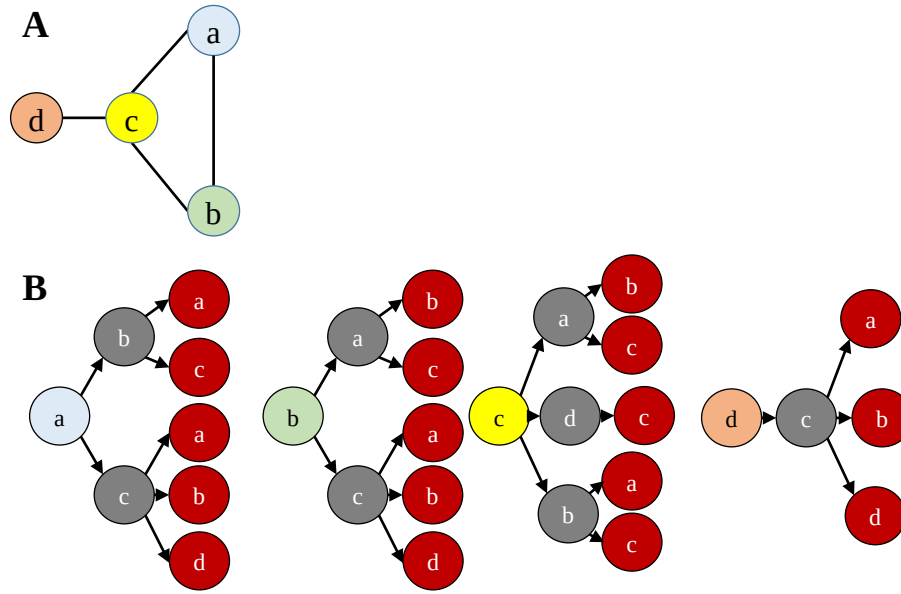# S.A. Ego-based and Alter-based Means in Example Networks

Fig. S1. Ego-based and Alter-based Means in Example Network. **(A) Network.** Example network with 4 nodes a,b,c and d. **(B) Illustration of Neighbors and Neighbors of Neighbors.** Each node is mapped out with its neighbors and neighbors of neighbors. The node is in light blue, neighbors are in gray, and neighbors of neighbors are in red. Node a has 2 neighbors, b and c. Node a also has 5 neighbors of neighbors.

Ego-based Mean: The ego-based mean number of neighbors of neighbors, detailed in (1), addresses the question, How many neighbors of neighbors does each node "experience" in the network? . Specifically, the first node a has two neighbors, and those two neighbors have $(2 + 3) = 5$ neighbors. Thus, on the average, ego a observes that her neighbors have $\frac{5}{2} = 2.5$ neighbors, which is more than she has herself. Similarly, ego b also has two neighbors, who in turn have $(2 + 3) = 5$ neighbors, for an average of $\frac{5}{2} = 2.5$ neighbors per neighbor. c on the other hand has three neighbors, and each neighbor has on the average only $\frac{2+1+2}{3} = \frac{5}{3}$ neighbors; thus, c's experience is that he has more neighbors than his neighbors do. Finally, ego d only has one neighbor, and that neighbor has 3 neighbors; thus, d's experience is that he has far fewer neighbors than his neighbors have. Overall, on the average, the four egos experience the following (ego-based) average number of neighbors that their neighbors have:

$$\mu_E = \frac{1}{\text{Number of nodes}} \sum_{i \in \text{Nodes}} (\text{Node } i\text{'s average number of neighbors of neighbors}) = \frac{1}{4}\left(\frac{5}{2} + \frac{5}{2} + \frac{5}{3} + \frac{3}{1}\right) = 2.42$$

Alter-based Mean: The alter-based mean, detailed in (2), addresses a similar idea but focuses on the average number of neighbors each alter has, where alters are the immediate neighbors of egos. For our purposes here, an alter is a special kind of neighbor; it is simply defined as an immediate neighbor of ego. In a network of N nodes, there are precisely N egos; but each node will frequently count more than once in their role as an alter, resulting in many more alters than egos in the network. That is, each node exists once in the network, but they play the role of alter perhaps several times (once for each of the nodes they are connected to). In the present example containing four egos, there are eight alters (colored gray in Panel B in the example above), and each alter has a certain number of adjacent neighbors. Counting how many neighbors each alter has is straightforward: The first alter above (b, who is an alter of a) has two neighbors (a and c); the second alter above (c, who is also an alter of a) has three neighbors (a, b, and d); and so on. Thus, the total number of neighbors of these alters (i.e., all the nodes in red in Panel B) is $(2 + 3 + 2 + 3 + 2 + 1 + 2 + 3) = 18$. The alter-based mean is then:

$$\mu_A = \frac{\text{Total number of neighbors of alters}}{\text{Total number of alters}} = \frac{18}{8} = 2.25$$
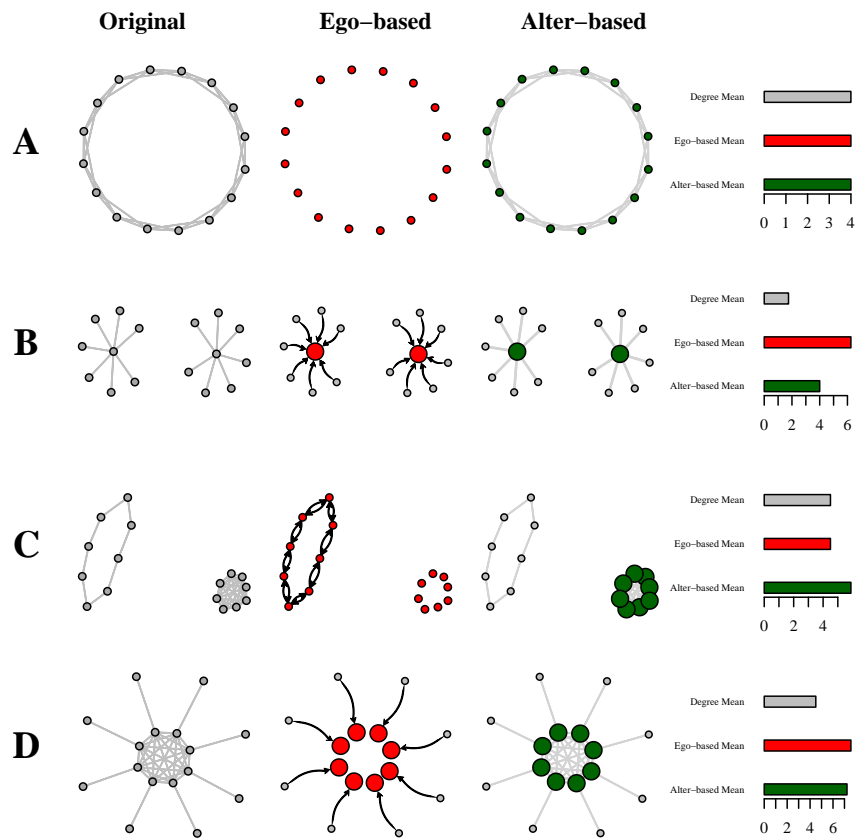
Vineet Kumar, David Krackhardt and Scott Feld

Fig. S2. Four Illustrative Networks with Varying Ego-based and Alter-based Means.
Each network in (A)-(D) has the original network plot (left), ego-based weighted network (middle), and alter-based weighted network (right). On the right is a barplot indicating the mean degree, ego-based mean, and alter-based mean for each of the networks. *Ego-based Panel (Red)*: In the weighted network plot (middle), nodes are sized proportional to their weight $\left(w_i^E\right)$ in contributing to the ego-based mean. Edges that receive a *higher than median* weight in computing the ego-based mean are in black. Otherwise, the edges are not plotted in the middle panel. Note that although the original networks are undirected, the selected edges are *directed*. *Alter-based Panel (Green)*: Nodes are sized proportional to their weight $\left(w_i^A\right)$ in contributing to the alter-based mean. Edges are all weighted equally in the alter-based weighted network. (A) *Small World Ring:* Each node has four friends, and ego-based and alter-based mean are both equal to the average degree (4). None of the edges are shown in the middle panel since all edges have identical weight in computing the ego-based mean. All nodes in both ego-based and alter-based means have the same weight and size in the middle and right panel. (B) *Two Central Hubs with Spokes:* Each central hub is connected to 7 nodes. The mean degree is lowest in this network. However, ego-based mean is substantially higher than the alter-based mean, and is higher than the mean degree across all networks (a)-(d). In the ego-based panel, we see that the weight of central hubs has increased, whereas the corresponding weight for the low degree "spoke" nodes has decreased. In the alter-based panel, the node weights are proportional to degree. (C) *Heavy Core with Detached Cycle:* The alter-based mean is substantially higher than the ego-based mean (and mean degree). Here, we see in the ego-based panel that the weight of each of the nodes has not changed, and all nodes have the same weight. However, in the alter-based panel, we see that the high degree nodes in the complete graph have higher weights compared to the original network, whereas the weights for the nodes in the 2-cycle are lower than in the original network. (D) *Heavy Core with Pendants:* Both the ego-based and alter-based mean are substantially higher than the mean degree. In the ego-based panel, the edges connecting core nodes to other nodes (both core and pendant) have a relatively low weight, and are not displayed.

S.B. Mathematical Appendix

Formally, the network graph $\mathcal{G} = (V, E)$ is comprised of a set of N individual nodes and a set of undirected edges E. Each
element of E is a pair of nodes, and (i, j) indicates an edge (connection) with $e_{ij} \in \{0, 1\}$. We also define the directed edge
set $\hat{E}$ including both (i, j) and (j, i) as distinct elements of $\hat{E}$ corresponding to an undirected edge $i \leftrightarrow j$. We use neighbor
and friend interchangably to try and connect with the literature. We note that neighbor is the more general terminology, and
appropriate in this article, since the network phenomena studied are not limited to social networks. We detail the table of
notation in Table S1.

Table S1. Table of Notation

Symbol	Term	Definition
$\mathcal{G}, V, E$	Network	Network Graph of Nodes V and Edges E
$\hat{E}$	Directed Edge Set	Each edge in E is replaced by two directed edges
$\mathcal{N}(i)$	Neighbors	*Set* of friends (neighbors) of i, $\mathcal{N}(i) = \{k \in V : \quad (i, k) \in E\}$
D_i	Degree	Number of friends (neighbors) of i, $D_i = |\{k \in V : \quad (i, k) \in E\}|$
F_i	Average degree of friends of i	$\frac{1}{D_i} \sum_{j \in \mathcal{N}(i)} D_j$
μ_D, σ_D^2	Mean and variance of Degrees	$\frac{1}{N} \sum_i D_i$, $\frac{1}{N} \sum_i (D_i - \mu_D)^2$
μ_E	Ego-based Mean	$\frac{1}{N} \sum_i F_i$
μ_A	Alter-based Mean	$\frac{\sum_i D_i F_i}{\sum_i D_i}$
ρ	Inversity	$Corr\left(D_i, \frac{1}{D_j}\right) \forall (i, j) \in \hat{E}$

The basic idea of the friendship paradox can be expressed as "your friends have more friends than you." We examine the
degree to which the friendship paradox holds for individual nodes, or the individual friendship paradox. We find in the result
below that it cannot hold for all nodes, but can hold for an arbitrarily high proportion (< 1) of nodes.

Theorem S1. *For a finite network $\mathcal{G} = (V, E)$ and $\mathcal{N}_i$ is the set of i's connections. We find the following:*

(i) The friendship paradox statement, "on average, your friends have more friends than you do," specified as $\frac{1}{|\mathcal{N}(i)|}\left(\sum_{j \in \mathcal{N}_i} D_j\right) >$
$D_i \ \forall i \in V$, cannot hold for all nodes in $\mathcal{G}$ or any connected component of $\mathcal{G}$.

(ii) There exists $\mathcal{G}$ for which the friendship paradox statement holds true for all nodes, except one.

Proof. Consider a multi-component network with C components, $V = \bigcup_{k=1}^{C} \mathcal{C}_k$, where each component $\mathcal{C}_k$ represents the set of
nodes in a connected network.
To prove part (i) of the theorem, first consider each of the components in turn, with $k = 1$. First, in the trivial case
of a degree-regular component, part (i) trivially holds. Next, consider the case with degree variation within component k.
Within C_k, for a finite network, there must be a finite set of nodes $\mathcal{V}_k^{\max}$ with maximum degree within this component. At
least one of the nodes in $\mathcal{V}_k^{\max}$ must then be connected to a node of lower degree; otherwise, the component would not be
fully connected. Now, for that node, call it $i \in \mathcal{V}_k^{\max}$ connected to a node of lower degree, the friendship paradox statement
$\frac{1}{|\mathcal{N}(i)|}\left(\sum_{j \in \mathcal{N}_i} D_j\right) > D_i \ \forall i \in \mathcal{C}_k$ cannot hold. Thus, for each component k, there is at least one node for which the friendship
paradox statement does not hold. In the overall network $\mathcal{G}$, there must be at least C nodes for which the friendship paradox
statement cannot hold.
For part (ii), we only need to consider the star or hub and spoke network. The friendship paradox statement can easily be
verified to hold for all nodes except the central node. $\square$

Theorem S2. *[Feld 1991] For a network $\mathcal{G} = (V, E)$ with degree mean μ_D and variance σ_D^2, the alter-based mean of friends of*
friends is $\mu_A = \left(\mu_D + \frac{\sigma_D^2}{\mu_D}\right)$

Proof. (as given in Feld, 1991). $\mu_A = \frac{\sum_i \sum_j e_{ij} D_j}{\sum_i D_i} = \frac{\sum_i D_i^2}{\sum_i D_i} = \frac{\mu_D^2 + \sigma_D^2}{\mu_D}$. We note that the above proof is not affected by
isolates, since they add zero to both the numerator and denominator, leaving μ_A unchanged, whether or not we remove these
isolates. $\square$

Theorem S3. *For any general network $\mathcal{G} = (V, E)$ with mean degree μ_D, the ego-based mean of friends is given by*

$$\mu_E = \mu_D + \frac{1}{2|V|} \sum_{(i,j) \in V \times V} e_{ij} \left[\frac{(D_i - D_j)^2}{D_i D_j}\right] \tag{1}$$

where D_i is the degree of node i, and $e_{ij} \in \{0, 1\}$ indicates a connection between i and j.

Vineet Kumar, David Krackhardt and Scott Feld

Proof. Let D_i denote the degree of i. Define $F_i = \frac{1}{D_i} \sum_{j \in N(i)} D_j$ as the mean number of neighbors for neighbors of i. The ego-based mean is defined as:

$$\mu_E = \frac{1}{|V|} \sum_i F_i = \sum_{i \in V} \left[\frac{1}{D_i} \left(\sum_{j \in \mathcal{N}(i)} D_j \right) \right]$$

Rewriting the expression for μ_E in terms of the connections (edges) between individuals, we obtain:

$$
\begin{aligned}
\mu_E &= \frac{1}{|V|} \sum_{i \in V} \left[\frac{1}{D_i} \left(\sum_{j \in V} e_{ij} D_j \right) \right] = \frac{1}{|V|} \sum_{i \in V} \sum_{j \in V} \left[e_{ij} \frac{1}{D_i} \left(D_j \right) \right] \\
&= \frac{1}{2|V|} \sum_{(i,j) \in V \times V} \left[e_{ij} \left(\frac{D_j}{D_i} \right) + e_{ji} \left(\frac{D_i}{D_j} \right) \right] = \frac{1}{2|V|} \sum_{(i,j) \in V \times V} e_{ij} \left[\frac{D_j}{D_i} + \frac{D_i}{D_j} \right] \\
&= \frac{1}{2|V|} \sum_{(i,j) \in V \times V} e_{ij} \left[\frac{D_j^2 + D_i^2}{D_i D_j} \right] = \frac{1}{2|V|} \sum_{(i,j) \in V \times V} e_{ij} \left[\frac{(D_i - D_j)^2 + 2D_i D_j}{D_i D_j} \right] \\
&= \frac{1}{2|V|} \sum_{(i,j) \in V \times V} e_{ij} \left[\frac{(D_i - D_j)^2}{D_i D_j} \right] + \frac{1}{2|V|} \left(4|E| \right) \\
&= \mu_D + \frac{1}{2|V|} \sum_{(i,j) \in V \times V} e_{ij} \left[\frac{(D_i - D_j)^2}{D_i D_j} \right]
\end{aligned}
$$

$\square$

Note that what we characterize as the ego-based mean defined above was independently shown to be greater than the mean degree, including by the present authors at (3). It has also been documented by others, including C. Borgs & J. Chayes in a later comment to an article by (4), and by (5). However, the properties of the ego-based mean have not been formally examined and characterized.

For the results below, we consider networks without isolates.

Theorem S4. *Define the m-th moment of the degree distribution by $\kappa_m = \frac{1}{N} \sum_{i \in V} D_i^m$. The ego-based and alter-based means are connected by the following relationship involving the inversity ρ and the -1,1,2, and 3rd moments of the degree distribution:*

$$\mu_E = \mu_A + \rho \sqrt{\left(\frac{\kappa_1 \kappa_3 - \kappa_2^2}{\kappa_1} \right) \left[\kappa_{-1} - (\kappa_1)^{-1} \right]}$$

Proof. We define the moments of the degree distribution as: $\kappa_m = \frac{1}{N} \sum_i D_i^m$. We defined inversity ρ as the correlation of two distributions that we specify as the origin degree (**O**) and inverse destination degree (**ID**) distributions. The **O** distribution consists of the degree of nodes corresponding to edges, and the **ID** distribution consists of the inverse degree of nodes corresponding to edges. Thus, each connection (edge) contributes *two* entries to *each* distribution. For example, if there is a connection between i and j, i.e., $e_{ij} = 1$, we would have $\left(D_i, \frac{1}{D_j} \right)$ and $\left(D_j, \frac{1}{D_i} \right)$. Observe that each individual appears in both distributions multiple times based on degree.

Next, we detail the mean and variance of the distributions. First, we consider the means. The mean of the origin distribution is $\mu_O = \frac{1}{2|E|} \sum_i D_i^2 = \frac{\mu_D^2 + \sigma_D^2}{\mu_D} = \mu_A = \frac{\kappa_2}{\kappa_1}$. Similarly, the **ID** mean is $\mu_{ID} = \frac{1}{2|E|} \sum_i D_i \left(\frac{1}{D_i} \right) = \frac{1}{\mu_D}$. Next, we consider the variances. The variance of the origin distribution (**O**) is computed as:

$$
\begin{aligned}
\sigma_O^2 &= \frac{1}{2|E|} \sum_{(i,j) \in E} (D_i - \mu_O)^2 = \frac{1}{2|E|} \sum_{i \in V} D_i (D_i - \mu_O)^2 \\
&= \frac{1}{N \mu_D} \sum_{i \in V} \left[D_i^3 - 2\mu_O D_i^2 + (\mu_O)^2 D_i \right] = \frac{\kappa_3}{\kappa_1} - \left(\frac{\kappa_2}{\kappa_1} \right)^2
\end{aligned}
$$

Next, we express the corresponding variance of the inverse destination degree distribution (**ID**), σ_{ID}^2. Again, recall that $\frac{1}{D_i}$ does not appear just once, but D_i times. Therefore, we have:

$$
\begin{aligned}
\sigma_{ID}^2 &= \frac{1}{2|E|} \sum_{(i,j) \in E} \left[\left(\frac{1}{D_j} - \frac{1}{\mu_D} \right)^2 \right] = \frac{1}{2|E|} \sum_{(i,j) \in E} \left(\frac{1}{D_j^2} + \frac{1}{\mu_D^2} - \frac{2}{\mu_D D_j} \right) \\
&= \frac{1}{2|E|} \left[\sum_{(i,j) \in E} \frac{1}{D_j^2} + \frac{1}{\mu_D^2} \left(\sum_{(i,j) \in E} 1 \right) - \frac{2}{\mu_D} \sum_{(i,j) \in E} \frac{1}{D_j} \right] = \frac{1}{2|E|} \left[\sum_{j \in V} \frac{1}{D_j} + \frac{1}{\mu_D^2} 2|E| - \frac{2}{\mu_D} N \right] \\
&= \frac{1}{\mu_D N} \left[\sum_{j \in V} \frac{1}{D_j} \right] - \frac{1}{\mu_D^2} \quad = (\kappa_1)^{-1} \left[\kappa_{-1} - (\kappa_1)^{-1} \right]
\end{aligned}
$$

We next turn to the inversity, and based on the definition, we connect it to the ego-based and alter-based means and the degree distribution.

$$\rho = \left(\frac{1}{2|E|\sigma_O\sigma_{ID}}\right) \sum_{(i,j)\in E} e_{ij}\left[(D_i - \mu_O)\left(\frac{1}{D_j} - \frac{1}{\mu_D}\right)\right]$$

$$(N\mu_D\sigma_O\sigma_{ID})\,\rho = \left[\sum_{(i,j)\in E} e_{ij}\left(\frac{D_i}{D_j}\right) - \mu_O\left(\sum_{(i,j)\in E}\frac{1}{D_j}\right) - \frac{1}{\mu_D}\sum_{(i,j)\in E} D_i + \sum_{(i,j)\in E} e_{ij}\left(\frac{\mu_O}{\mu_D}\right)\right]$$

$$= \left[N\left(\mu_E\right) - \mu_O\cdot N - \frac{1}{\mu_D}\sum_{(i,j)\in E} D_i + \sum_{(i,j)\in E} e_{ij}\left(\frac{\mu_O}{\mu_D}\right)\right]$$

$$= \left[(N\mu_E) - N\mu_O - \frac{1}{\mu_D}\sum_{(i,j)\in E} D_i + 2|E|\left(\frac{\mu_O}{\mu_D}\right)\right]$$

$$\implies \mu_E = \mu_A + \rho\cdot\mu_D\cdot\sigma_O\sigma_{ID}$$

Finally, substituting $\mu_D = \kappa_1$ and the expressions for the variances, we obtain:

$$\mu_E = \mu_A + \rho\sqrt{\left(\frac{\kappa_1\kappa_3 - \kappa_2^2}{\kappa_1}\right)\left[\kappa_{-1} - (\kappa_1)^{-1}\right]} \qquad [2]$$

$\square$

Theorem S5. *The expected degree of nodes chosen by alter-based strategy is the alter-based mean.*

Proof. To determine the expected degree of a node chosen by the alter-based strategy: Choose $M = 1$ node initially, (say X). With probability q, choose each neighbor of X. For a node k with degree D_k, the probability of being chosen by this process is the first step when any of k's neighbors is chosen as the initial node, and the second step is k being chosen with probability q. This probability is $p_k = \frac{1}{N}\,D_k \times q = \frac{qD_k}{N}$. The expected degree of a chosen "seed" node is then the degree-weighted probability:

$$\frac{\sum_{k\in V}\ p_k D_k}{\sum_{k\in V} p_k} = \frac{\sum_{k\in V}\frac{1}{N}q\ D_k^2}{\sum_{k\in V}\frac{1}{N}q\ D_k} = \frac{\frac{1}{N}\sum_{k\in V} D_k^2}{\frac{1}{N}\sum_{k\in V} D_k} = \frac{\mu_D^2 + \sigma_D^2}{\mu_D} = \mu_A$$

$\square$

Similar logic applies if we choose any arbitrary initial sample of size M as long as the network is large, i.e., $N \gg M$.

Theorem S6. *[Rewiring Theorem] Let network $\mathcal{G} = (V, E)$ with $N > 3$ nodes include nodes a, b, c, d with degrees ordered as: $D_a \leq D_b < D_c \leq D_d$. If there are nodes $a, b, c, d \in V$ such that $(a,b), (c,d) \in E$, but $(a,d), (b,c) \notin E$, then by rewiring the network to $\mathcal{G}' = (V, E')$, containing edges $(a,d), (b,c) \in E'$, but $(a,b), (c,d) \notin E'$, we obtain: $\mu_E(\mathcal{G}') > \mu_E(\mathcal{G})$ whereas $\mu_A(\mathcal{G}') = \mu_A(\mathcal{G})$. Also, it follows that $\rho(\mathcal{G}') > \rho(\mathcal{G})$.*

Proof. First, observe that the degree distribution is unaffected by the change, since each node's degree is unchanged by the rewiring. Therefore, the alter-based mean (which only depends on mean and variance of the degree distribution) is also unaffected, i.e., $\mu_A(\mathcal{G}) = \mu_A(\mathcal{G}')$. Recall that the ego-based mean is $\mu_E = \frac{1}{N}\sum_i\sum_j e_{ij}\left[\frac{D_i}{D_j} + \frac{D_j}{D_i}\right]$. Since between $\mathcal{G}$ and $\mathcal{G}'$ the degrees of all nodes are the same, and all edges are the same except the two rewired edges, we can write the difference between their ego-based means as:

$$\mu_E(\mathcal{G}') - \mu_E(\mathcal{G}) = \frac{1}{N}\left[\left(\frac{D_a}{D_d} + \frac{D_d}{D_a} + \frac{D_b}{D_c} + \frac{D_c}{D_b}\right) - \left(\frac{D_a}{D_b} + \frac{D_b}{D_a} + \frac{D_c}{D_d} + \frac{D_d}{D_c}\right)\right]$$

$$= \frac{1}{N}\left[(D_d - D_b)\left(\frac{1}{D_a} - \frac{1}{D_c}\right) + (D_c - D_a)\left(\frac{1}{D_b} - \frac{1}{D_d}\right)\right] > 0$$

The last inequality follows from the ordering of the node degrees. Note that we actually only require the conditions $D_b < D_d$ and $D_a < D_c$ to hold. Since the degree distribution does not change with rewiring, by Theorem S4, we must have an increase in inversity, $\rho(\mathcal{G}') > \rho(\mathcal{G})$.

$\square$

S.C. Data on Real Networks

We use a wide variety of real networks to characterize their properties, and illustrate how relate to the interventions detailed in the paper. We use data from two repositories.

A. Koblenz Network Collection. The networks are selected across several categories (Affiliation, Face-to-face Social, Online Social, Computer, Infrastructure and Biological networks), and span a wide range in network characteristics like size and density (Table S2). These networks also vary widely in terms of their size, from a low of 25 to networks with millions of nodes (e.g., Youtube). All network data was obtained from the Koblenz Network Collection (KONECT) (6). We examine these real networks on a number of dimensions, including the number of nodes, edges, and the variation in the degree distribution.

Table S2. Real Network Characteristics

Label	Network	N	$|E|$	μ_D	μ_E	μ_A	ρ
Collaboration							
A1	Actor-Movie	383640	1470338	7.67	36.29	35.12	0.02
A2	Club Mmebers	25	91	7.20	10.39	9.39	0.39
A3	Citation (Physics)	28045	3148413	224.53	569.15	667.24	-0.06
A4	Citation (CS)	317080	1049865	6.62	18.53	21.75	-0.10
Face-to-Face Interaction							
FS1	Physician	117	464	7.93	10.19	9.95	0.09
FS2	Adolescent Health	2539	10454	8.23	9.85	10.49	-0.20
FS3	Contact	274	2124	15.50	74.78	56.69	0.26
FS4	Conference	410	2765	13.49	17.10	18.72	-0.19
Online Social							
OS1	PGP Users	10679	24315	4.55	13.46	18.88	-0.17
OS2	Flickr	105722	2316667	43.83	187.12	349.21	-0.22
OS3	Advogato	5042	40509	15.56	99.31	82.52	0.06
OS4	Twitter	465016	833539	3.58	437.74	226.53	0.65
Topology of Computer Networks							
C1	Internet Topology	34761	107719	6.20	530.34	319.46	0.22
C2	WWW (Google)	855802	4291352	10.03	226.59	170.35	0.06
C3	Gnutella P2P	62561	147877	4.73	13.22	11.60	0.15
Infrastructure							
I1	Power Grid	4941	6593	2.67	3.97	3.87	0.06
I2	US Airports	1572	17214	21.90	120.27	112.23	0.04
I3	CA Roads	1957027	2760387	2.82	3.15	3.17	-0.04
Biological							
B1	Human Protein 1	2783	6222	4.32	19.61	15.78	0.15
B2	Human Protein 2	5973	146385	48.81	117.83	143.31	-0.09
B3	Yeast Protein	1458	1970	2.67	9.65	7.13	0.31
B4	C. Elegans	453	2033	8.94	51.57	40.10	0.21

A few observations are worth noting here. First, there are many networks with both positive and negative values of inversity, both within and across categories. Second, we do not see Inversity ρ close to ± 1. However, the Twitter network is closest in magnitude, with an inversity of $\rho = 0.65$. Third, the variation in inversity is low in some categories like Infrastructure, whereas it is relatively greater in Online Social networks. Finally, we see that even low values of inversity can impact the difference between the ego-based and alter-based means substantially, as well as between each of these and the mean degree. The WWW (Google) network, for example, displays such meaningful differences, even with a low inversity value of 0.06. This is due to the multiplier effect of the moments of the distribution function, detailed in equation (3).

B. India Village Networks. In addition, we also use data from $N = 75$ villages in India made publicly available (see (7) for details). The summary statistics for those village household networks are detailed in Table S3.

Table S3. Summary Statistics of Village Networks

Network Statistic	Mean	SD	Min	Max
Number of households	216.69	61.22	77	356
Number of (undirected) edges	993.31	348.77	334	2015
Density	0.05	0.02	0.02	0.11
Degree Mean	9.10	1.573	6.13	12.78
Degree Variance	52.03	19.88	27.80	124.56

S.D. Individual Friendship Paradox

A basic view of the friendship paradox is developed by plotting the average number of friends (degree) of individual nodes'
"friends" on the vertical axis against the average degree (Fig. S3, Fig. S4). For example, in the Contact (In-person Social)
network, we see a deep blue region above and to the left of the 45° line. Although present across all networks, the pattern is
most prominent in the WWW (Google) or Twitter (Online Social) network. Observe also that in the Road Network, only
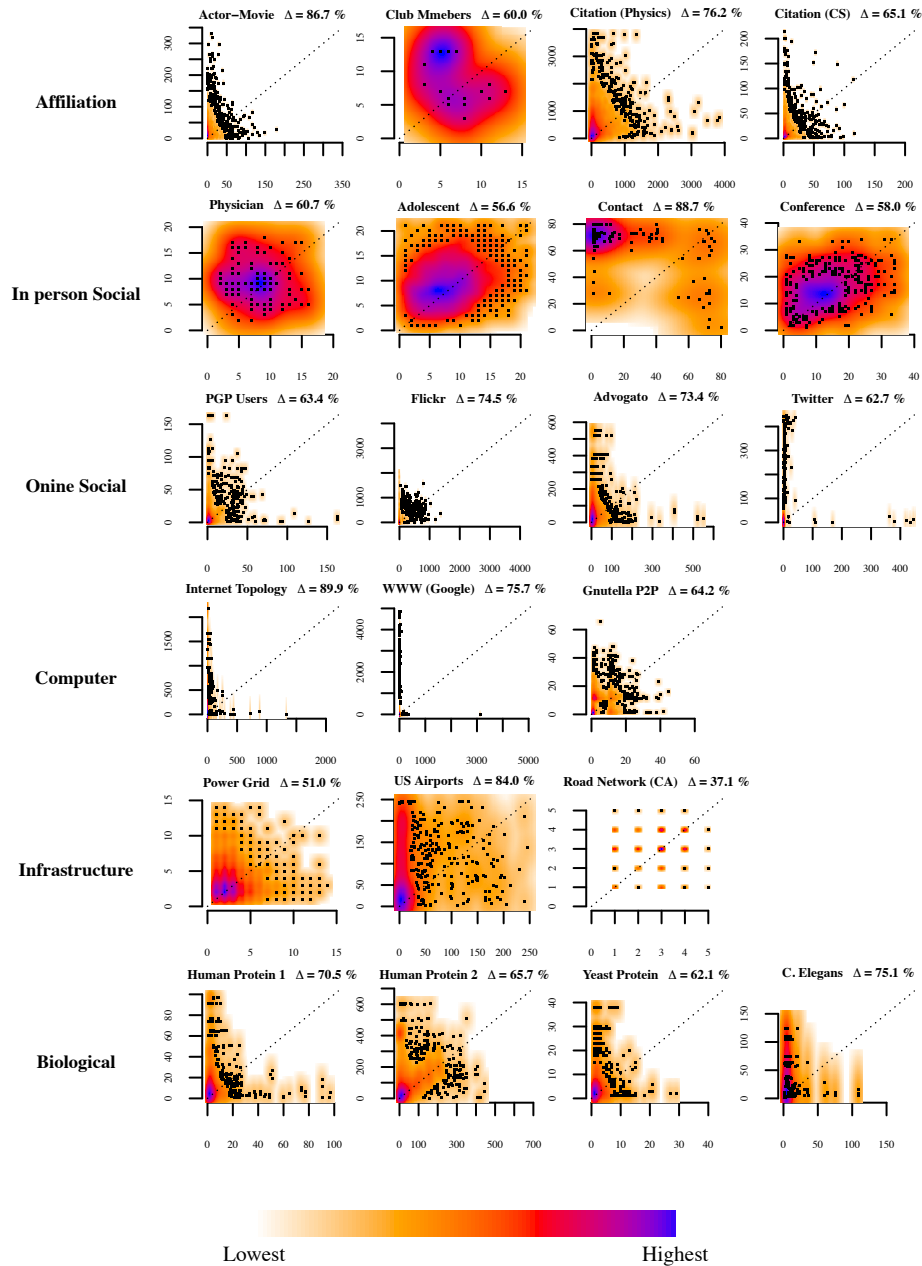$\Delta = 37\%$ of nodes have a higher average number of friends of friends than their own degree.

Fig. S3. Friendship Paradox at Individual Level. Density plot of average number of friends of nodes compared to node degree in networks. Δ indicates the proportion of nodes that have a higher average number of friends of friends than their degree. Lowest density regions within each network are marked by white / orange, and highest density regions are marked in blue. For all networks, the highest density region lies above and to the left of the 45 degree line. For some networks like Adolescent Health or Road Network (CA), it is relatively more evenly distributed both above and below the 45 degree line, whereas for networks like Internet Topology or Twitter, the distribution is skewed above and to the left.
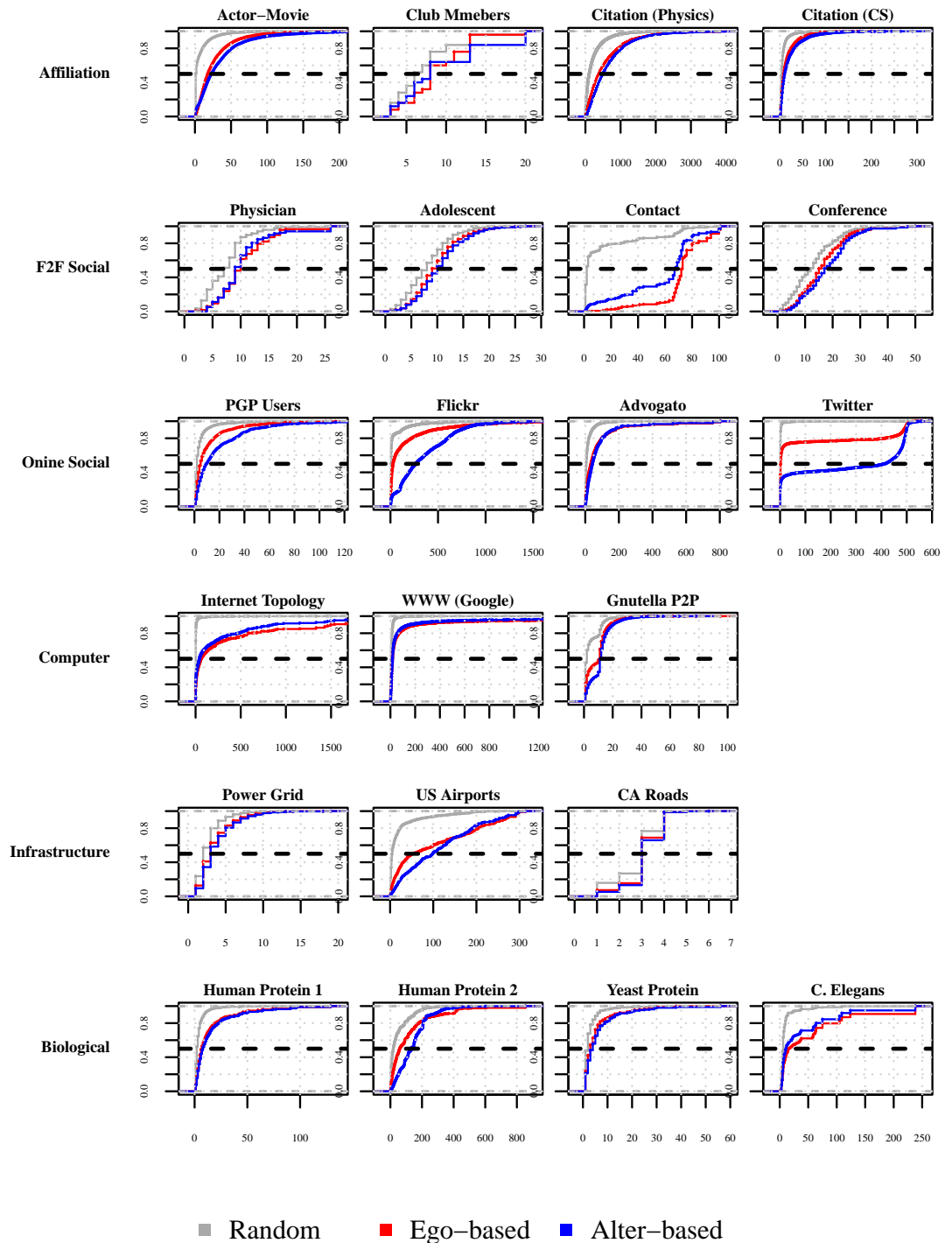
Fig. S4. Individual Friendship Paradox. Empirical Cumulative Distribution Functions (CDF) of Real Networks. Panels show the CDF of 3 different network properties at the individual node level. For a specific node degree, the probability that a node with a lower (or identical) degree is chosen by the sampling strategy for node degree (gray), ego-based mean of Friends of Friends (in red), and alter-based mean of friends of friends (in blue). Across all networks, for lower degrees, the node degree curve is to the left of the ego-based and alter-based mean of friends curves. In several networks, alter-based mean of friends is to the left and higher than ego-based mean of friends (e.g., Contact), whereas in others, it is to the right (e.g., Flickr).
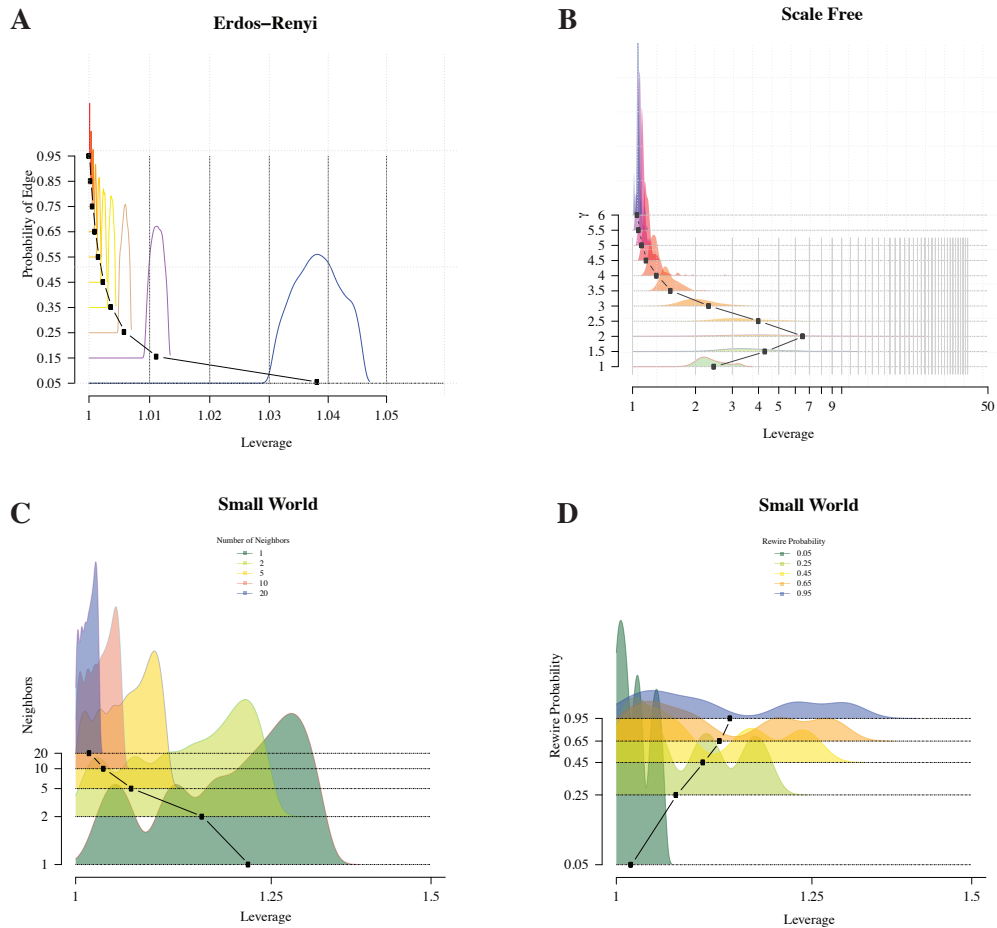
S.E. Leverage in Generated Networks

Fig. S5. Ego-based Leverage Density in Generated Networks from three different generative models, spanning the parameter space of each. A sample of 1,000 networks was used for each of the models. (A) Erdos-Renyi (ER) networks generated with edge probabilities, $p \in [0.05, 0.95]$, and size ranging from N=50 to N=1,000 nodes. We find that ego-based leverage is highest for the lowest edge probabilities, and leverage converges to 1 as the networks become more dense. (B) (SF) Static Scale Free networks with scale-free parameter $\gamma \in [1, 6]$ (8). For these networks, observe that the leverage spans a wider range, e.g., for $\gamma = 2$, the samples range from leverage of 1 to over 40. The mean leverage is non-monotonic in terms of γ, increasing when $\gamma < 2$ and decreasing for $\gamma > 2$. The distribution of leverage across the samples also displays decreasing variance when $\gamma > 2$. At very high levels of $\gamma \approx 6$, the ego-based mean converges to the mean degree. With (SW) small world networks, we have two parameters (9). First is the number of neighbors each node is connected to initially, n. The edges are then rewired with a specified probability, p_r. First, in panel (C), we find that with a small number of neighbors, the leverage distribution is quite spread out, and there is a substantial leverage effect. However, as we begin to create very dense networks, both the mean and the variance of the leverage distribution diminish substantially. Second, we examine the impact of rewiring probability on the leverage distribution in panel (D). We find that with lower rewiring probabilities, say $p_r = 0.05$, the leverage distribution is closer to 1, whereas with higher rewiring probabilities, the distributions feature increased variance as well as higher mean leverage.

S.F. Network Features that Impact Inversity

Inversity is strongly dependent on the structure of connections, who is connected to whom. We observe in the main paper that star-type network structures lead to positive values of inversity, whereas clusters or cliques contribute to negative inversity values.

Star or hub-based networks   First, we observe that there is significant evidence for hub-based network structures appearing in real-world networks. Such hub-like structures are common across a wide range of networks, including co-author networks, the underlying interlink network that forms the Internet, as well as Airline networks (10–13). The early networks literature explaining the emergence of such hub-like patterns posited preferential attachment as a mechanism, where newly joining nodes connected disproportionately to highly connected nodes (8, 14). The economics literature involving the economic incentives underlying network formation posits that agents form links based on the expected benefits to such formation. The resulting network is an equilibrium outcome based on the decisions of each of the agents, who are maximizing their own utilities (15). In this stream, an influential paper (16) finds that even when agents are homogeneous — where they have identical constraints, preferences, and incentives — *star networks arise across a wide range of equilibria.* Stars are predicted to occur even though agents are symmetric with identical incentives and opportunities. Complementing this research, star networks are found to arise over time in experimental settings where agents vary in terms of costs, incentives, and even information (17).

Clusters and Communities in networks   Clusters or communities as well as cliques (fully connected or complete subnetworks) are commonly observed in networks. A typical conceptualization of community is the following: "Qualitatively, a community is defined as a subset of nodes within the graph such that connections between the nodes are denser than connections with the rest of the network." (18). There are several reasons why communities form, including homophily and social foci. In homophily, when a number of individuals are similar, then they are much more likely to be connected to each other, and also part of the same larger grouping or community (19). However, it should be noted that not all such connections will happen; rather, such connections and communities are more likely to happen when individuals are homophilous. We note that homophily has also been tied to polarization and segregation (20).

A prominent theory that explains how communities form is the idea of foci (21). The essential idea is that most ties originate around foci of activity, where a limited set of people share a focus that organizes activity, and thereby tend to generate repeated interaction among the same people in the set over time that leads to ties among many of them. Each person tends to be associated with many different foci. Alters from the same focus tend to be tied to one another, but not those from separate foci. Consistent with this notion, research has found that the way organizational environments are structured moderates the tie-formation process (22).

An implication of this theory for the present paper is that larger, denser foci of activity contribute large numbers of ties for all their tied participants, and small and/or sparse focused sets generate few for their participants. Thus, the size and density of focused sets may contribute to positive or negative inversity.

S.G. Inversity and Assortativity: Connections and Differences

A natural question is whether inversity, $\rho = Corr\left(D^{\mathbf{O}}, \frac{1}{D^{\mathbf{D}}}\right)$, captures the same information (with opposite sign) as degree assortativity, which is a well known network property, $\rho_a = Corr\left(D^{\mathbf{O}}, D^{\mathbf{D}}\right)$ (23–25). Inversity and assortativity are negatively correlated, as we might expect.

Can Assortativity Be Used as a Proxy for Inversity?. There are several specific reasons why we don't think it is a good idea to use assortativity as a *proxy metric* in place of inversity. We demonstrate specifically how using assortativity as a proxy for inversity would lead to incorrect choices to the network intervention questions below.

Network Intervention Questions: We begin with our objective for the network intervention strategies, which is to choose a strategy that maximizes the expected degree of target nodes. We have two questions that need to be answered.

(a) Identify whether the ego-based strategy leads to higher expected degree than the alter-based strategy (or vice versa), i.e. whether $\mu_E > \mu_A$ or $\mu_A > \mu_E$.

(b) Evaluate the improvement in expected degree offered between any of random, ego-based and alter-based strategies, i.e. $(\mu_E - \mu_D)$, $(\mu_A - \mu_D)$ and $(\mu_E - \mu_A)$.

The requirements for problem (a) are different from those of (b). For (a), we need to just know the correct ordering of ego-based and alter-based strategies, and not the magnitude. Inversity (ρ) gives us a direct answer to question (a), since $\mu_E > \mu_A \iff \rho > 0$. For (b), the ordering is not sufficient, and we need to know the magnitude, and we explore this below.

Why Magnitudes of Differences Matter: The decision maker could evaluate the benefit of using a friendship paradox strategy is worth the potential cost relative to the random strategy, which requires the least amount of information and effort from the people originally selected. Thus, the decision maker would trade off the increase in expected degree relative to the marginal cost of using the strategy (26). This logic implies that knowing just the ordering of the different strategies is not sufficient, and *we would need to know the magnitudes of the differences* in order to select a strategy.

Inversity allows for a direct linear transformation from alter-based mean (μ_A) to ego-based mean (μ_E) calculation of neighbors of neighbors. We can see this using the formulation from equation (2) in page 5 of this Supplement:

$$\mu_E = \mu_A + \rho \underbrace{\left(\sqrt{\left(\frac{\kappa_1 \kappa_3 - \kappa_2^2}{\kappa_1}\right)\left[\kappa_{-1} - (\kappa_1)^{-1}\right]}\right)}_{\omega} = \mu_A + \rho\,\omega$$

where the κs represent the moments of the degree distribution. This formula implies that $\mu_E = \mu_A + \omega\rho \implies \mu_E - \mu_A = \omega\rho$, where ω is a function of the moments of the degree distribution.

We have found no comparable transformation between alter-based and ego-based means using assortativity, and argue that such a relationship is unlikely to exist (see point below about monotonic ordering between these metrics). Thus, to use assortativity in place of inversity, we must first assume $\rho_a \approx -\rho$ in order to make an approximation of the form: $\mu_E - \mu_A \approx -\omega\rho_a$. We don't know whether and when this approximation would be valid.*

If we do make this approximation, high values of the quantity ω can amplify small differences in the values of ρ. Therefore, even *small errors* in approximating inversity with assortativity would be highly problematic in evaluating magnitude of effect sizes. We also see that this impact is practically important. In Table S2 of Supplement §S.C, for instance, we observe that even small values of inversity can result in large differences between the different means. Please see the statistics corresponding to the A3 (Citation) and C2 (WWW) networks.

Vineet Kumar, David Krackhardt and Scott Feld

Decisions using Assortativity versus Inversity: Inversity is sensitive to and positively affected by stars and star-like structures. Assortativity, on the other hand, is more sensitive to and positively affected by cliques and clique-like structures. Supplement §S.F shows that cliques and stars are commonly present in real world networks.

Illustrative Example Networks by Simulation: We explore this logic quantitatively by simulating specific type of networks with stars and cliques to make things concrete. We examine how inversity and assortativity change as the size of the star network is varied, keeping the cliques constant. To make this deterministic, we have fixed the cliques to include all cliques from size 2 to size 20. We add only one star network and vary the size of the star network, with the number of nodes in the star varying from 1 to 61.

We plot in Figure S6 both the assortativity and inversity levels for each network as a function of how large the star is in the network. All other parameters remain fixed. The blue circles indicate the inversity (ρ) for the generated network; the red plus-signs indicate the assortativity (ρ_a) for the same generated network. The plot also includes inset sociograms (A)-(D) at specific points of interest.

First, we note that inversity and assortativity are highly correlated, with $cor(\rho, \rho_a) = -0.992$. Similarly, we test a linear regression model of ρ and ρ_a and find the R^2 of the model to be $R^2 = 0.983$. Thus, the measures seem on the surface to be very closely related. However, we observe if we use assortativity as a proxy for inversity, we would not make the right decisions for questions (a) or (b). Using assortativity as proxy, we would approximate $\rho \approx -\rho_a$.

We now detail the problems with using assortativity:

(1) In the shaded region, where the size of the star networks is between 29 and 47 (approximately), we observe that the sign of assortativity and inversity is the same, i.e., *both are positive.* In such a case, the answers to both questions (a) and (b) would be incorrect.

(2) To the left of the shaded region, where the size of the star network varies from 2 to 29 (approximately), we note that $|\rho_a| > |\rho|$, implying that even though the sign of ρ_a and ρ are opposite, the magnitudes are quite different. More specifically, assuming $\rho \approx -\rho_a$, we would overestimate the benefit of the ego-based strategy, leading to errors in answering question (b).

(3) To the right of the shaded region, i.e., when star networks have size greater than 47, we observe that $|\rho| \gg |\rho_a|$. This difference in magnitudes implies that using the proxy assumption $\rho \approx -\rho_a$ would be highly problematic for obtaining magnitudes required for (b). More specifically, we would substantially underestimate the benefits of the ego-based strategy relative to alter-based and random.

Broadly, in these networks, assortativity cannot reliably help us determine which mean (ego-based or alter-based) is larger, or which strategy dominates. We also don't know the conditions under which it could serve (or not) as a reasonable proxy for inversity for these or more general networks. Thus, even when these metrics are highly correlated, using the proxy of assortativity in place of inversity could lead to errors over a substantial range of networks.

Monotonic Ordering between Assortativity and Inversity?   Since Assortativity (ρ_a) as a proxy is directionally the *reverse* of Inversity (ρ), we would naturally expect that it to vary monotonically with inversity. Specifically, we would expect for any pair of networks $\mathcal{G}_1$ and $\mathcal{G}_2$, when $\rho(\mathcal{G}_1) > \rho(\mathcal{G}_2) \iff \rho_a(\mathcal{G}_1) < \rho_a(\mathcal{G}_2)$. However, there are (many) examples of pairs of networks where both assortativity are directionally the same, i.e. $\rho(\mathcal{G}_1) > \rho(\mathcal{G}_2)$ and $\rho_a(\mathcal{G}_1) > \rho_a(\mathcal{G}_2)$. Using common network generation algorithms, it is easy to obtain many such pairs of networks, where one of the networks has both **higher assortativity *and* higher inversity** than the other. Thus, we find that assortativity as a proxy metric for inversity does not even preserve ordering, making it problematic to rely on it as a proxy for network intervention problems. When we examined inversity and assortativity for a set of 45 village networks from India, we found 3 of the 45 village networks had both positive inversity and assortativity.

Variation of Inversity and Assortativity with Set of Cliques. [†] We also experimented with varying sets of cliques to understand how and when inversity and assortativity diverge. First, we begin with all cliques of size 2 to 15 included in a network. We then alter the network to selectively remove cliques. The sociograms indicate the cliques that are included in the network.

[†] We thank an anonymous reviewer for suggesting this exercise.

Vineet Kumar, David Krackhardt and Scott Feld

A

Size of star = 2
Cliques with sizes
ranging from 2 to 20

Inversity = −0.708

Assortativity = 1

B

Size of star = 30
Cliques with sizes
ranging from 2 to 20

Inversity = 0.027

Assortativity = 0.553

Inversity ρ and Assortativity ρ_a

−1.0   −0.5   0.0   0.5   1.0

cor(ρ, ρ_a)=−0.992

0

10

20

30

40

50

60

Size of star

○ Inversity
+ Assortativity

C

Size of star = 39
Cliques with sizes
ranging from 2 to 20

Inversity = 0.296

Assortativity = 0.228

D

Size of star = 47
Cliques with sizes
ranging from 2 to 20

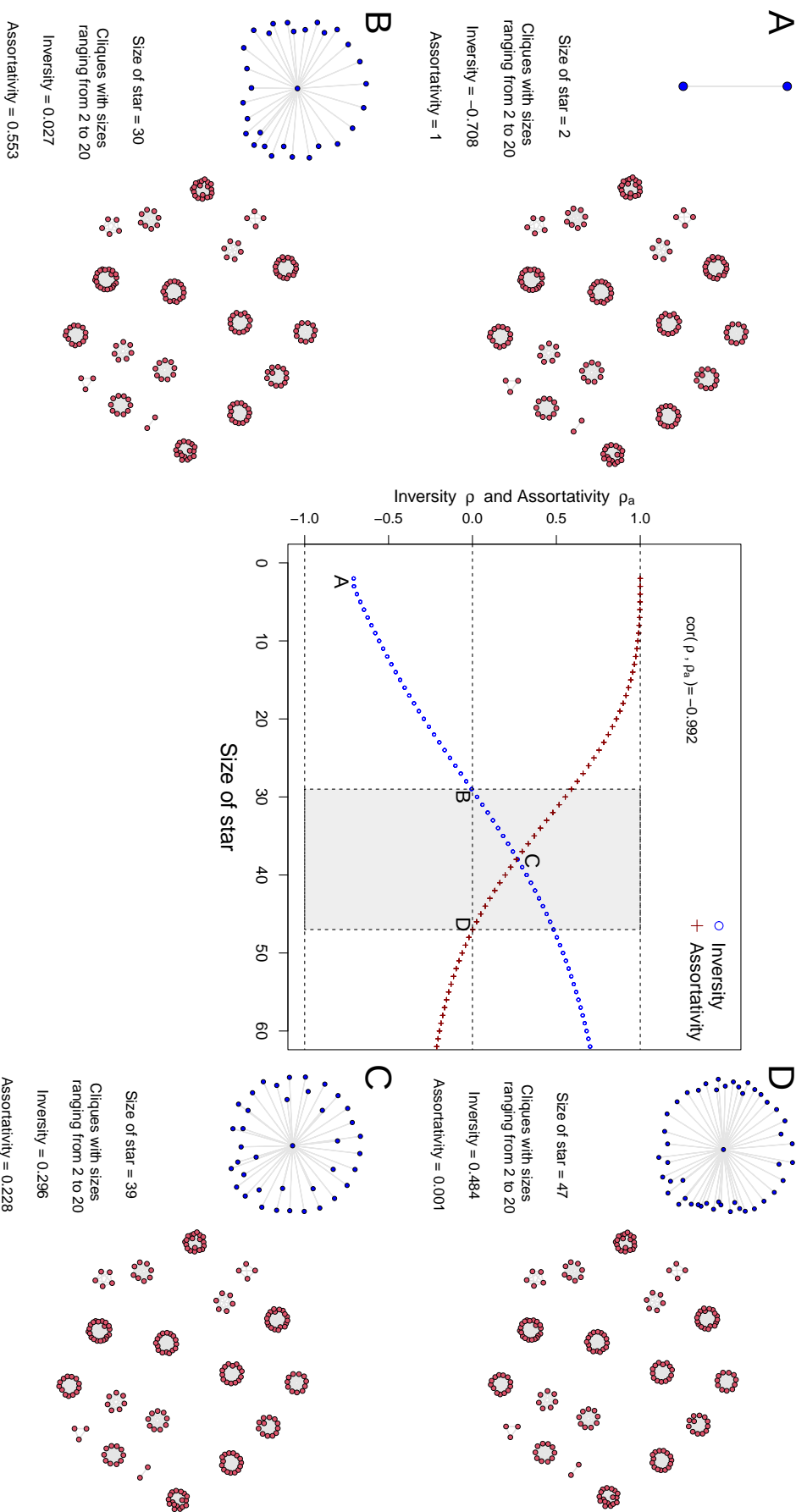Inversity = 0.484

Assortativity = 0.001

Fig. S6. Variation of Inversity with Star Size

Note: We simulate networks with one star and several cliques. The size of the cliques ranges from 2 to 20 nodes. The size of the star network is varied between 2 and 60 (one hub connected to 60 spokes). The set of cliques is maintained exactly the same across all of the networks. Each network appears in the plot as a point. We observe that inversity is increasing with size of the star network, whereas assortativity is decreasing. The shaded area represents a region among these networks where *both inversity and assortativity are positive*. We illustrate the network topologies corresponding to 4 different networks (A)–(D), where (A) corresponds to the a star of size 2, (B) has a star of size 30, (C) has a star of size 39, and (D) has a star of size 47.

Each panel corresponds to a network, which here is a set of cliques, ranging in size from 2 to 15. We illustrate the degrees of nodes connected by edges in each panel, with each edge in the network corresponding to a point in the plot. The points (and lines) in blue corresponding to (d_i, d_j) depicting edge (i, j) that connects nodes i and j. This illustration views the degrees the way assortativity would see them. The points (and lines) in red plots $(d_i, 1/d_j)$, corresponding to the way inversity would view the edge (i, j).
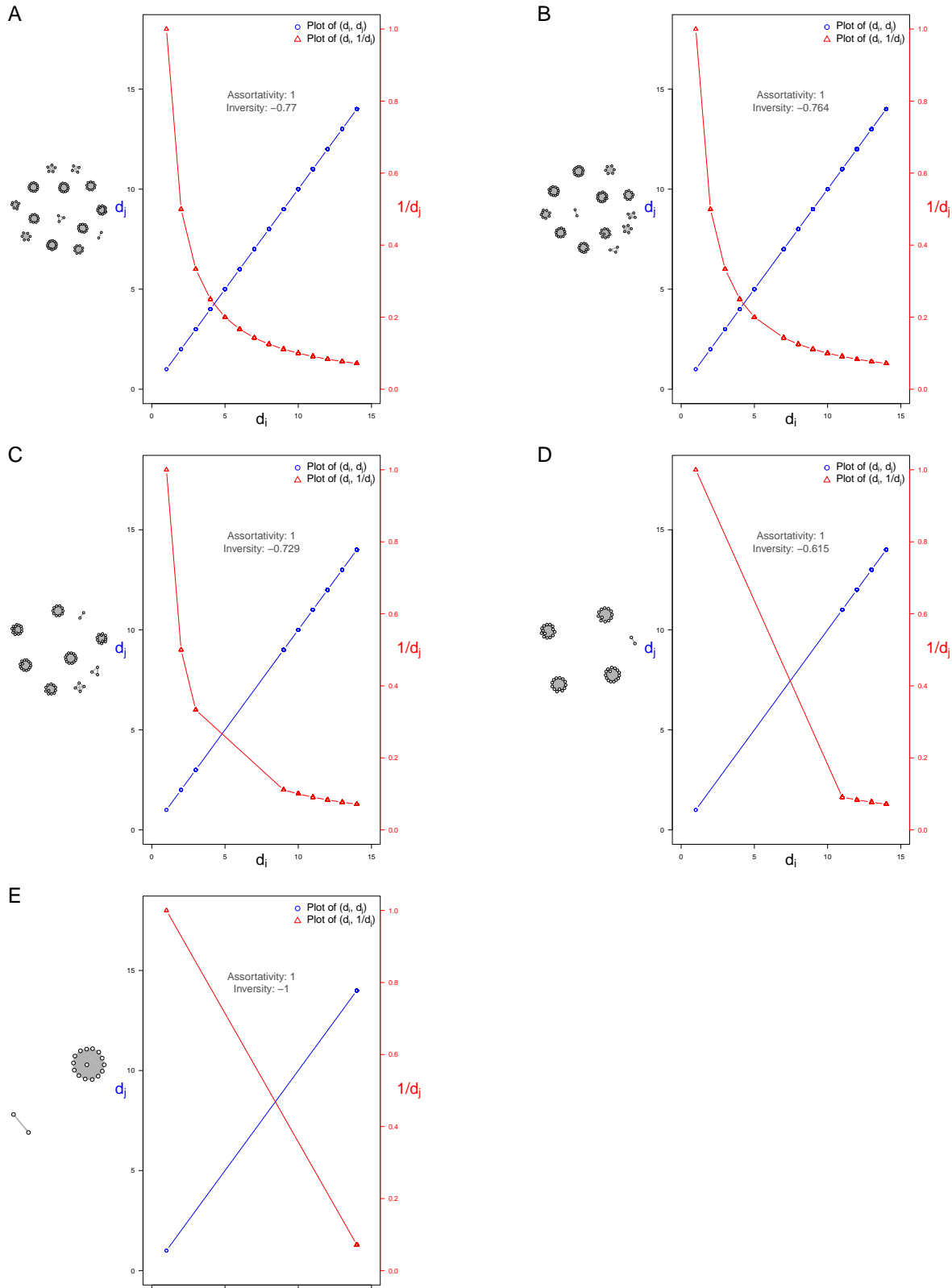
We then plot the degrees of nodes that connect an edge in the x-axis and y-axis. For edge (i, j), we include (d_i, d_j) in the plot in blue. For the same edge, we also include $(d_i, 1/d_j)$ in red. Thus, we observe a few different things.

1. We observe that assortativity for all the sets of cliques we have tested is always 1. However, the inversity is only -1 in the trivial case where we have 2 cliques.

2. Inversity and assortativity have a non-linear relationship with the degree as expected. Specifically, when we have low degree nodes, they contribute disproportionately more to inversity.

3. The inversity across the set of cliques appears to vary in a non-monotonic manner. Specifically, we have $\rho = -0.77$ in panel (A), where we have all the cliques between size 2 and 15 included in the network. As we remove cliques, we find that in panel (B), the inversity $\rho = -0.764$ is closer to zero. This pattern continues in (C) and (D) when we remove even more cliques, we observe that inversity increases and becomes less negative. However, in the extreme case when we have only 2 cliques (size 2 and size 15), we observe that inversity $\rho = -1$. Thus, we see that there does not appear to be a simple pattern for how inversity varies with the number or set of cliques. We think future research might find it useful to examine the patterns of variation of inversity, and how it contrasts with assortativity.

S.H. Virus Propagation Models

We detail below several examples of virus propagation models being used for characterizing the transmission and spread of diseases. These models build upon the early work of Kermack and McKendrick (27). All individuals in a population (in our case, the nodes in a network) are in one of the states, either susceptible (S) or infected (I). Based on the viral propagation, they can move to other states like Exposed (X), Recovered (R), or Deceased (D). For example, the SIR model involves individuals being in one of three states, (S), (I) or (R), and transitioning between the states probabilistically. Typically, the vast majority of nodes are present in the susceptible state (S), in which they might contrast the disease. The exposed state (X) is used to indicate a node that has been exposed to the disease, but could be asymptomatic during an incubation period and is not capable of infecting others. In contrast, the infected state (I) indicates a node that is capable of infecting others. The (R) recovered state implies permanent immunity. There are further extensions possible, e.g., adding infants who have maternal antibodies (state M) that provide passive immunity. See (28) or (29) for an overview and survey of these models. These models have been extensively used in epidemiological studies to characterize disease dynamics as detailed in Table S4, including measles, influenza, and COVID-19.

There has been recent notable work that aims to characterize the epidemic thresholds of these compartmental models with disease transmission over a network (30, 31). The critical idea is that the epidemic threshold of a network can be characterized as the inverse of the greatest (first) eigenvalue of the adjacency matrix A of the network, denoted as:

$$\tau(A) = \frac{1}{\lambda_1(E)}$$

.

Eigenvalue λ_1, termed the spectral radius, characterizes the connectivity of the network graph. Thus, networks that have higher connectivity or λ_1 are more likely to allow contagions along different paths to grow into epidemics, whereas in networks with low connectivity, contagions are more likely to die out.

While there have been a number of epidemic thresholds for specific network generating processes (e.g., small world), the generality of the result above is valuable since it allows: (a) any arbitrary network, without placing restrictions on its topology or structure, and (b) a wide range of compartmental models like SIS, SIR, and others detailed in Table S4 typically used to model infectious disease.

Whereas we consider an SIR model for illustration, the results also hold for the other models. The model is parametrized by two rates: β is the probability of an infected node infecting a susceptible node in a given time period, and δ is the probability at which an infected node recovers (or is cured) during the period. If time is continuous, β and δ can be viewed as the rates of infection and recovery. In either case, $\mathcal{R}_0$ is defined as $\boxed{\mathcal{R}_0 = \frac{\beta}{\delta}}$.

The epidemic threshold τ is defined as follows (30):

$$\begin{cases} \mathcal{R}_0 = \frac{\beta}{\delta} < \tau(E) \implies \text{infection dies out over time} \\ \mathcal{R}_0 = \frac{\beta}{\delta} > \tau(E) \implies \text{infection grows over time} \end{cases}$$

There are a few observations relevant here. First, the critical value of epidemic threshold is a function of the adjacency matrix E of the network topology (structure) $\mathcal{G}$. Second, a network topology with a higher epidemic threshold is less likely to have an epidemic. Third, interventions like immunizing nodes or reducing the number of connections (edges) can increase the threshold $\tau(E)$ so that infections are more likely to die out.

Table S4. Virus Propagation Models Used for Diseases

Virus Propagation Model	Infectious Diseases [References]
SIS	Malaria (32)
SIR	Measles (33), Swine Flu H1N1 (34), Ebola (35)
SXIR	Chicken Pox (36), SARS (37), COVID-19 (38)
SIRD	COVID-19 ((39))

Note: The states refer to (**S**)usceptible, (**I**)nfectious, (**R**)ecovered / (**R**)emoved, (**X**)Exposed, (**D**)eceased

Implementation of VPM. We begin with a seed set of 1% of the nodes being infected, and evaluate epidemic outcomes using the SIR model. All the nodes in the network that are not infected or recovered are susceptible (S) to the infection. Each infected node can transmit an infection in each period probabilistically to each of its neighbors. The probability of an infection is $P_{\text{transmit}} = \beta$. Thus, a node can become infected (I) from contact with any of its neighbors. In each period, an infected node can be cured or recovered (R) probabilistically, with the likelihood $P_{\text{cure}} = \delta$. Recovered nodes cannot be reinfected and cannot transmit infections.

The process of immunizing (or vaccinating) a set of nodes involves choosing a proportion of nodes (5%, or 10% or 20%) and ensuring that these nodes do not transmit any disease. The nodes for immunization are chosen based on three strategies: random, ego-based, and alter-based. The parameters used in the simulation of the epidemic are detailed in Table S5.

Table S5. Parameters of SIR Network Propagation Model

Parameter	Value	Description
$P_{\text{transmit}} = \beta$	0.20	Probability of an infected node transmitting the disease to a susceptible neighbor.
$P_{\text{cure}} = \delta$	0.15	Probability of an infected node recovering. Thus, moving from (I) $\implies$ (R) is $P_{I \to R} = P_{\text{cure}}$, and $P_{I \to I} = 1 - P_{\text{cure}}$
$P_{S \to I}^k$	$1 - (1-\beta)^{N_k^{\text{infected}}}$	Probability of a susceptible node k becoming infected. Depends on the number of infected neighbors N_k^{infected}. Thus, k can become infected through *any* of its infected neighbors. So we have: $P_{S \to I}^k = 1 - (1 - P_{\text{transmit}})^{N_k^{\text{infected}}}$. Similarly, $P_{S \to S}^k = (1 - P_{\text{transmit}})^{N_k^{\text{infected}}}$.
n_{infected}^0	1%	Proportion of nodes in network that are infected at the beginning
n_{sim}	100	Number of simulations

Note: (**S**)usceptible, (**I**)nfectious, (**R**)ecovered / (**R**)emoved

Thus, a strategy A is better than an alternative strategy B if it results in lower levels of peak infections, total infections, and total suffering.

S.I. Epidemic Outcomes

In Figure S8, we examine the epidemic propagation characteristics on the Facebook network (40) using the same parameters as detailed in Table S5. The epidemic could be viewed as an *informational* epidemic propagating through Facebook. Alternatively, one might consider the Facebook network structure to serve as an approximation of contact network for the purposes of this evaluation.

We evaluate epidemics using the following metrics:

- **Proportion Infected at Peak** $= \frac{1}{N} \max_t(\sum_i I_{it})$: Since epidemics increase in intensity and eventually die down, an important characteristic is to measure the proportion of the population who are infected at the peak of the epidemic. This directly impacts important decisions like hospital capacity planning, etc.

- **Proportion Ever Infected** $= \frac{1}{N} \sum_i \max_t(I_{it})$: The proportion of the population that was ever infected by the disease is important since it represents the total spread of the disease in the population. It could also represent the number of people who might have immunity to future recurrences of the disease.

- **Total Suffering**: $\frac{1}{NT}\sum_i\sum_t(I_{it})$ Here, the total suffering metric captures not just how many infections occur, but also the length of the infections. This represents the proportion of individual-period combinations with an infection.

For the Facebook network sample, we find that an epidemic's outcomes are better when using the ego-based strategy compared to the alter-based strategy, which in turn is better than the random strategy. This conclusion holds for all the metrics considered above.
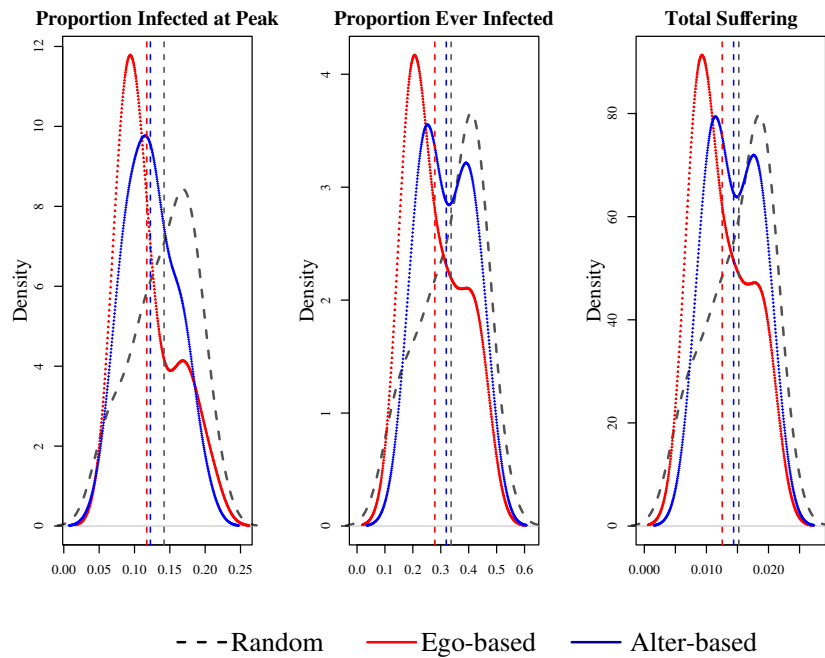

Fig. S8. Epidemic Outcomes with Immunization in Facebook Network. See Table S5 for parameters of simulation. All outcomes are density plots. We plot 3 outcomes: (a) the proportion of population infected at the peak, (b) proportion of population that was ever infected, and (c) total suffering. In each panel, the x-axis represents proportions and the y-axis represents density. We plot the outcomes for 3 strategies: (R)andom, (E)go-based, and (A)lter-based. The dashed vertical lines represent the means for the 3 strategies. We find that for the Facebook network, the ego-based strategy is better for all outcomes than the alter-based strategy, which in turn is better than the random strategy.

Vineet Kumar, David Krackhardt and Scott Feld

References

1. D Krackhardt, Structural leverage in marketing. *Networks marketing* pp. 50–59 (1996).
2. S Feld, Why your friends have more friends than you do. *Am. J. Sociol.* pp. 1464–1477 (1991).
3. V Kumar, D Krackhardt, S Feld, Why your friends are more popular than you are - a revisit in *Archives of the 33rd Sunbelt Conference of International Network for Social Network Analysis.* (INSNA), (2013).
4. S Strogatz, Friends you can count on. *The New York Times* **September 17** (2012).
5. MO Jackson, The friendship paradox and systematic biases in perceptions and social norms. *J. Polit. Econ.* **127**, 777–818 (2019).
6. J Kunegis, Konect: the Koblenz network collection in *Proceedings of the 22nd International Conference on World Wide Web.* (ACM), pp. 1343–1350 (2013).
7. A Banerjee, AG Chandrasekhar, E Duflo, MO Jackson, The diffusion of microfinance. *Science* **341**, 1236498 (2013).
8. AL Barabási, R Albert, Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
9. D Watts, S Strogatz, Collective dynamics of "small-world" networks. *nature* **393**, 440–442 (1998).
10. AL Barabâsi, et al., Evolution of the social network of scientific collaborations. *Phys. A: Stat. mechanics its applications* **311**, 590–614 (2002).
11. ME Newman, Coauthorship networks and patterns of scientific collaboration. *Proc. national academy sciences* **101**, 5200–5205 (2004).
12. S Goyal, MJ Van Der Leij, JL Moraga-González, Economics: An emerging small world. *J. political economy* **114**, 403–412 (2006).
13. ME O'Kelly, HJ Miller, The hub network design problem: a review and synthesis. *J. Transp. Geogr.* **2**, 31–40 (1994).
14. AL Barabási, , et al., Scale-free networks: a decade and beyond. *science* **325**, 412 (2009).
15. MO Jackson, , et al., *Social and economic networks.* (Princeton university press Princeton) Vol. 3, (2008).
16. V Bala, S Goyal, A noncooperative model of network formation. *Econometrica* **68**, 1181–1229 (2000).
17. JK Goeree, A Riedl, A Ule, In search of stars: Network formation among heterogeneous agents. *Games Econ. Behav.* **67**, 445–466 (2009).
18. F Radicchi, C Castellano, F Cecconi, V Loreto, D Parisi, Defining and identifying communities in networks. *Proc. national academy sciences* **101**, 2658–2663 (2004).
19. M McPherson, L Smith-Lovin, JM Cook, Birds of a feather: Homophily in social networks. *Annu. review sociology* **27**, 415–444 (2001).
20. P Dandekar, A Goel, DT Lee, Biased assimilation, homophily, and the dynamics of polarization. *Proc. Natl. Acad. Sci.* **110**, 5791–5796 (2013).
21. SL Feld, The focused organization of social ties. *Am. journal sociology* **86**, 1015–1035 (1981).
22. DA McFarland, J Moody, D Diehl, JA Smith, RJ Thomas, Network ecology and adolescent social structure. *Am. sociological review* **79**, 1088–1121 (2014).
23. ME Newman, Assortative mixing in networks. *Phys. review letters* **89**, 208701 (2002).
24. ME Newman, J Park, Why social networks are different from other types of networks. *Phys. Rev. E* **68**, 036122 (2003).
25. I Sendiña-Nadal, MM Danziger, Z Wang, S Havlin, S Boccaletti, Assortativity and leadership emerge from anti-preferential attachment in heterogeneous networks. *Sci. reports* **6**, 21297 (2016).
26. D Krackhardt, Structural leverage in marketing in *Networks in marketing*, ed. D Iacobucci. (Sage, Thousand Oaks), pp. 50–59 (1996).
27. A M'Kendrick, Applications of mathematics to medical problems. *Proc. Edinb. Math. Soc.* **44**, 98–130 (1925).
28. F Brauer, The Kermack–McKendrick epidemic model revisited. *Math. biosciences* **198**, 119–131 (2005).
29. HW Hethcote, The mathematics of infectious diseases. *SIAM review* **42**, 599–653 (2000).
30. D Chakrabarti, Y Wang, C Wang, J Leskovec, C Faloutsos, Epidemic thresholds in real networks. *ACM Transactions on Inf. Syst. Secur. (TISSEC)* **10**, 1–26 (2008).
31. BA Prakash, D Chakrabarti, M Faloutsos, N Valler, C Faloutsos, Got the flu (or mumps)? check the eigenvalue! *arXiv preprint arXiv:1004.0060* (2010).
32. D Smith, J Dushoff, R Snow, S Hay, The entomological inoculation rate and plasmodium falciparum infection in african children. *Nature* **438**, 492–495 (2005).
33. MJ Ferrari, et al., The dynamics of measles in sub-saharan africa. *Nature* **451**, 679–684 (2008).
34. M Small, C Tse, Clustering model for transmission of the SARS virus: application to epidemic control and risk assessment. *Phys. A: Stat. Mech. its Appl.* **351**, 499–511 (2005).
35. T Berge, JS Lubuma, G Moremedi, N Morris, R Kondera-Shava, A simple mathematical model for Ebola in africa. *J. biological dynamics* **11**, 42–74 (2017).
36. S Deguen, G Thomas, NP Chau, Estimation of the contact rate in a seasonal SEIR model: application to chickenpox incidence in france. *Stat. medicine* **19**, 1207–1216 (2000).
37. S Riley, et al., Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science* **300**, 1961–1966 (2003).
38. K Prem, et al., The effect of control strategies to reduce social mixing on outcomes of the covid-19 epidemic in Wuhan, China: a modelling study. *The Lancet Public Heal.* (2020).
39. D Caccavo, Chinese and Italian covid-19 outbreaks can be correctly described by a modified sird model. *medRxiv* (2020).

40. B Viswanath, A Mislove, M Cha, KP Gummadi, On the evolution of user interaction in Facebook in *Proc. Workshop on Online Social Networks.* pp. 37–42 (2009).

Vineet Kumar, David Krackhardt and Scott Feld