Supporting Information

Ugander et al. 10.1073/pnas.1116502109

SI Text

Supporting Analysis of Recruitment. ***Five-node neighborhood topologies.*** As an extension of Fig. 1 *B–D*, we present the recruitment rates for invitation neighborhoods consisting of five nodes (Fig. S1). We note that when studying neighborhoods with more than five nodes, the recorded data are spread thinly across an overwhelming number of possible graph topologies, and considering every topology is no longer possible.

Structural vs. demographic diversity. As a potential confounder for our findings, we consider the fact that neighborhoods with many components are comparatively likely to also exhibit increased demographic diversity, which may figure into conversion in a manner outside our structural analysis. To control for this, to the extent that it is possible, we condition our data on neighborhoods that are demographically homogenous with respect to self-reported sex, geography, and age, meaning that all of the site users within the neighborhood are of the same sex, from the same country, and all contained within a 5-year range of age. We note that for neighborhoods >2 in size, this homogeneity requirement entails an aggressive restriction on the amount of admissible data, to the point that for neighborhoods composed of a four-node cycle, we observe no converted registrations. We find that the significance of our structural measure of neighborhood diversity persists in this demographically controlled dataset (Fig. S2).

Embeddedness and weak ties. Here we study the role of Granovetter's structural measure of weak and strong ties, termed *embeddedness*. For an individual $i \in V$ in a social graph $G = (V, E)$, let his or her neighborhood graph N_i be the subgraph of G induced by his or her neighboring nodes $V_i = \{j \in V : e_{ij} \in E\}$. *Weak ties*, in a structural sense, are ties with low embeddedness in the social graph, where the embeddedness of edge e_{ij} is the number of common neighbors of the two node endpoints, $\mathrm{Em}(e_{ij}) = |N_i \cap N_j|$. As an equivalent definition, the embeddedness of edge e_{ij} is also equal to the degree of node v_j within the neighborhood of node v_i, $\mathrm{Em}(e_{ij}) = \deg_{N_i}(j)$. Through this, we observe that the *embeddedness distribution* of a neighborhood, $\mathrm{Em}(N_i) = \{\mathrm{Em}(e_{ij}) : j \in N_i\}$, is the same as the degree distribution of the neighborhood.

Granovetter's work on "the strength of weak ties" found that unembedded edges—those with embeddedness zero, termed *local bridges*—play an important role in the spread of awareness for new opportunities, specifically in the labor market (1). Applying this principle of information novelty to our recruitment domain suggests that invitations arriving along edges with low embeddedness may be more likely to result in successful recruitment. As a consequence, if i is a node who accepts an invitation, one might expect that at least some neighbors j of i will tend to be connected via edges e_{ij} of low embeddedness. In other words, the embeddedness distribution $\mathrm{Em}(N_i)$ will have small values: Viewed as a multiset, it will contain small numbers as elements.

This observation leads to a potential confounding effect in our analysis of connected components, in the following way. As noted above, the embeddedness distribution of the neighborhood N_i is the same as its degree distribution; hence, small values in this distribution are consistent with a sparse structure for N_i and hence with the potential for N_i to have many components. What if the relationship between the number of components and the probability of recruitment is in fact a consequence of the relationship between small numbers in the embeddedness distribution and the probability of recruitment?

Fortunately, we can separate these effects quite cleanly, as follows. There exist pairs of graphs on five and six nodes with precisely the same degree distribution, but with different numbers of connected components (Fig. S3*A*). If we look for invitees i whose contact neighborhoods come from these pairs, we will have neighborhoods whose degree distributions—and hence whose embeddedness distributions—are identical, but have different numbers of connected components. Any argument based on embeddedness values has no way to distinguish among these pairs of graphs and hence would necessarily predict equivalent rates of recruitment.

Analyzing recruitment rates on precisely these five- and six-node topologies pushes the resolution limits of what is possible even with huge amounts of data, but even so we see that for every such pair of graphs in which the embeddedness distributions are identical but the component counts differ, the neighborhood with more components has a higher rate of recruitment (Fig. S3*B*). Thus, diversity, measured by component count, appears to play an important role in recruitment conversion in a manner decidedly outside traditional theories of information diffusion.

***k*-Braces.** In this section we present formal results regarding the notion of a k-brace defined in this work. Recall from the main text that the k-brace is constructed by repeatedly deleting edges of embeddedness less than k until there are no such edges remaining, followed by a single pass in which all isolated nodes are deleted. The first thing we prove is that this procedure leads to a well-defined outcome. Indeed, some iterative update procedures of this general flavor can potentially produce different end results depending on the order in which the updates are performed; what we wish to show is that the final subgraph produced by the k-brace procedure does not in fact depend on the order in which the edge deletions are performed. To do this, we provide a succinct graph-theoretic characterization of this final subgraph and then show that all ways of scheduling the edge deletions lead to this subgraph. Finally, we give an efficient algorithm, adapted from the work of Cohen (2), for computing the k-brace.

To characterize the end result of the edge deletion process, we begin with the following definition. Given a graph $G = (V, E)$, a subgraph H of G is a pair (W, F), where $W \subseteq V$ and $F \subseteq E$, and each edge in F has both endpoints in W. We now define the following collection of subgraphs $\mathcal{B}_k(G)$: We say that a subgraph H of G belongs to $\mathcal{B}_k(G)$ if (*i*) each edge of H belongs to at least k distinct triangles in H and (*ii*) each node of H has at least one incident edge in H. We observe that $\mathcal{B}_k(G)$ is a nonempty set, because the subgraph consisting of no nodes and no edges satisfies conditions *i* and *ii* and hence belongs to $\mathcal{B}_k(G)$.

To motivate the definition of $\mathcal{B}_k(G)$, note that the outcome of the procedure defining the k-brace of G produces a subgraph in $\mathcal{B}_k(G)$. We wish to show more, namely that the k-brace is in fact the unique maximal element of $\mathcal{B}_k(G)$ under a certain natural partial order. In particular, for two subgraphs of G, denoted $H_1 = (W_1, F_1)$ and $H_2 = (W_2, F_2)$, we say that $H_1 \preceq H_2$ if $W_1 \subseteq W_2$ and $F_1 \subseteq F_2$. We now claim.

Proposition 1. *In the set $\mathcal{B}_k(G)$, partially ordered by $\preceq$, there is a unique maximal element.*

Proof. Let us define the following union operation on subgraphs: If $H_1 = (W_1, F_1), H_2 = (W_2, F_2), \ldots, H_s = (W_s, F_s)$ are subgraphs of G, then we define their union $\cup_{i=1}^{s} H_i$ to be the subgraph $(\cup_{i=1}^{s} W_i, \cup_{i=1}^{s} F_i)$.

The key fact underlying the proof is that if $H_1 = (W_1, F_1), H_2 = (W_2, F_2), \ldots, H_s = (W_s, F_s)$ are subgraphs in $\mathcal{B}_k(G)$, then $\cup_{i=1}^{s} H_i$ also belongs to $\mathcal{B}_k(G)$. To see why, we simply observe that

(i) every edge in $\cup_{i=1}^{s} H_i$ belongs to at least one of the H_i and hence is part of at least k triangles and (ii) every node in $\cup_{i=1}^{s} H_i$ belongs to at least one of the H_i and hence is incident to at least one edge.

Given this result, if we enumerate all of the subgraphs $H_1, H_2, \ldots, H_t$ in $\mathcal{B}_k(G)$, then their union $\cup_{i=1}^{t} H_i$ is also an element of $\mathcal{B}_k(G)$. It is the unique maximal element of $\mathcal{B}_k(G)$, because for any subgraph H in $\mathcal{B}_k(G)$, the subgraph H is one of the elements in the union $\cup_{i=1}^{t} H_i$, and hence $H \preceq \cup_{i=1}^{t} H_i$. ∎

Let $\beta_k(G)$ denote the unique maximal element of $\mathcal{B}_k(G)$. We now claim the following.

Proposition 2. *Any execution of the procedure defining the k-brace, regardless of the order of edge deletions, results in the subgraph* $\beta_k(G)$.

Proof. Consider an execution of the edge deletion procedure, removing edges in the order $e_1, e_2, \ldots, e_s$. Let $G_j = (V, E - \{e_1, e_2, \ldots, e_{j-1}\})$ be the subgraph of G after the first $j-1$ edge deletions, at the moment just before e_j was deleted.

We claim that none of the deleted edges $e_1, e_2, \ldots, e_s$ belong to any subgraph in $\mathcal{B}_k(G)$. Indeed, suppose by way of contradiction that this were not the case, and consider the first edge e_j that does belong to a subgraph $H = (W, F)$ in $\mathcal{B}_k(G)$. In $H = (W, F)$, the edge e_j belongs to a set of k distinct triangles; let $T \subseteq F$ be the set of $2k$ edges other than e_j that constitute these triangles. None of the edges $e_1, e_2, \ldots, e_{j-1}$ can belong to T, because by assumption e_j is the first edge in the sequence of deletions to belong to any subgraph in $\mathcal{B}_k(G)$. However, this observation implies that all of the edges of T were still present in the underlying graph G_j at the moment that e_j was considered for deletion, and because e_j therefore belonged to at least k distinct triangles in G_j, it should not have been deleted—a contradiction.

Similarly, we claim that none of the isolated nodes deleted at the end of the procedure belong to any subgraph in $\mathcal{B}_k(G)$. Again, suppose by way of contradiction that one of the deleted nodes v belonged to a subgraph $H = (W, F)$ in $\mathcal{B}_k(G)$. In H, node v is incident to some edge e. However, e was not present when v was deleted, and hence e itself must have been deleted earlier in the procedure; hence, $e \in \{e_1, e_2, \ldots, e_s\}$. However, we have just shown that none of the edges in $\{e_1, e_2, \ldots, e_s\}$ belong to any subgraph in $\mathcal{B}_k(G)$, whereas e belongs to H, a contradiction.

Finally, consider any execution of the edge deletion procedure, and let H^* denote the subgraph that results from it. H^* belongs to $\mathcal{B}_k(G)$, because at the termination of the procedure all edges in H^* have embeddedness at least k and there are no isolated nodes, and hence by Proposition 1, $H^* \preceq \beta_k(G)$. On the other hand, we have just established that any node or edge that belongs to any subgraph in $\mathcal{B}_k(G)$ also belongs to H^*, and hence $\beta_k(G) \preceq H^*$. It follows that $H^* = \beta_k(G)$, as desired. ∎

Finally, we describe the following efficient implementation of the edge deletion procedure for computing the k-brace, adapted from Cohen (2).

Algorithm 1: Extracting the k-brace. *Given a graph G and a parameter k, use a queue q to efficiently traverse the graph and iteratively remove all edges with embeddedness <k.*

```
for e ∈ G.edges() do
  Em(e) ← size(G.neighbors(e [0]) ∩ G.neighbors(e [1]));
  if Em(e) < k then
    q.append(e);
    G.removeEdge(e);
  end if
end for
while size(q) != 0 do
  e← q.pop();
  I ← G.neighbors(e [0]) ∩ G.neighbors(e [1])
  for v ∈ I do
    e′ ← (e [0],v);
    Em(e′) ← Em(e′) − 1;
    if Em(e′) < k then
      q.append(e′);
      G.removeEdge(e′);
    end if
    e″ ← (e [1],v);
    Em(e″) ← Em(e″) − 1;
    if Em(e″) < k then
      q.append(e″);
      G.removeEdge(e″);
    end if
  end for
end while
for v in G.nodes() do
  if degree(v) == 0 then
    G.removeNode(v);
  end if
end for
```

For a graph with n nodes and m edges, a straightforward analysis shows that the runtime for this algorithm is at most $O(\sum_{v \in V} \text{degree}^2(v)) = O(m^2)$, which is rather expensive, but fortunately our focus on neighborhood graphs implies that all of the graphs we consider are very modest in size.

As mentioned in the text, the k-brace is always a subgraph of the $(k+1)$-core. Because finding the $(k+1)$-core takes merely $O(n+m)$, in practice it is more efficient to first compute the $(k+1)$-core of a graph G and then find the k-brace of the $(k+1)$-core rather than the full graph G; the analog of this optimization is also present in Cohen's work (2).

Supporting Analysis of Engagement. ***Predicted engagement for other neighborhood sizes.*** Here we present results that extend Fig. 4 *D–F* from the main text (Fig. S4). Similar to our comments in the main text, note that when comparing neighborhoods of different sizes, we can see that having a 30-node neighborhood with two components in the 1-brace predicts as much engagement as having a 50-node neighborhood with only one component in the 1-brace. ***Controlling for k-brace size.*** Fig. S5 presents a control of the potential confounding factor that k-braces with multiple components may have a tendency to be larger. Here we control for the size of the 1-brace to show that predicted engagement is still an increasing function of component count when controlling for size.

1. Granovetter M (1973) The strength of weak ties. *Am J Sociol* 78:1360–1380.

2. Cohen JD (2008) Trusses: Cohesive subgraphs for social network analysis. *National Security Agency Technical Report* (National Security Agency, Fort Meade, MD).

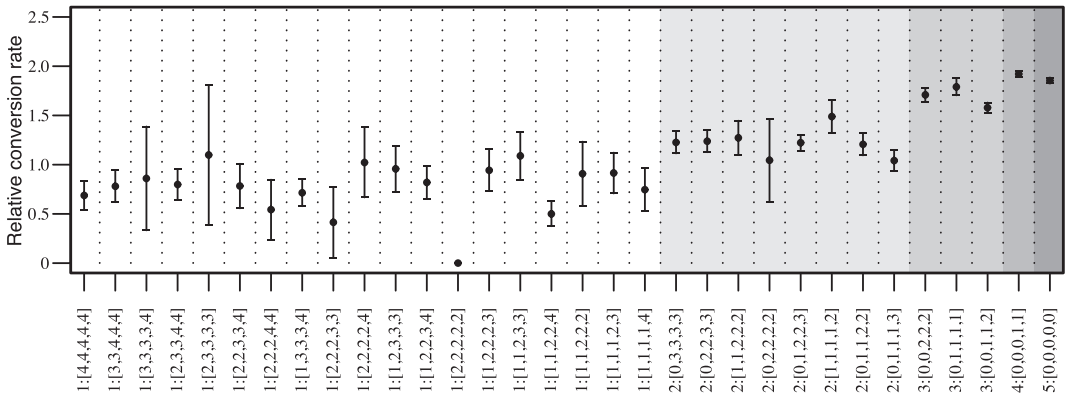

Fig. S1. Invitation conversion rates of size five neighborhoods, labeled by their number of connected components and their degree distribution, which uniquely determines the topology (for graphs of size 4 or smaller, degree distribution is a unique determinant of component count, but this uniqueness is not true for graphs of size 5 or larger). For example, the label "1:[4,4,4,4,4]" indicates the clique, 1 component where all of the nodes have degree 4. The label "2:[1,1,2,2,2]" indicates a graph of two components (a triangle and a pair), whereas the label "1:[1,1,2,2,2]" indicates the one-component line graph. The five-cycle topology "1:[2,2,2,2,2]" was exceedingly rare, and no conversions for this topology were observed. The conversion scale is the same as for Fig. 1 in the main text. Error bars are 95% confidence intervals.

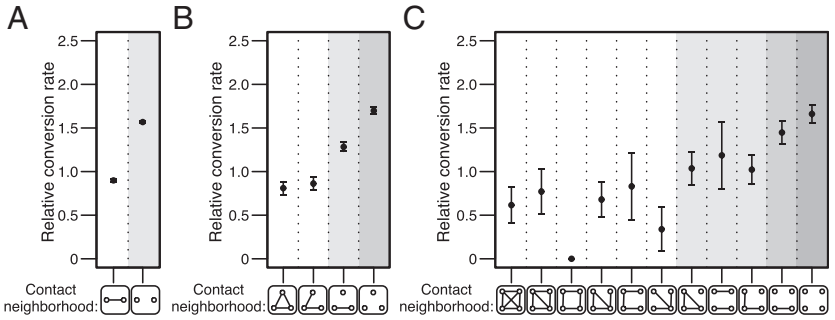

Fig. S2. Recruitment conversion for demographically homogeneous neighborhoods, as a function of (*A*) two-node, (*B*) three-node, and (*C*) four-node contact neighborhood graphs. The conversion scale is the same as for Fig. 1 in the main text. Error bars represent 95% confidence intervals.

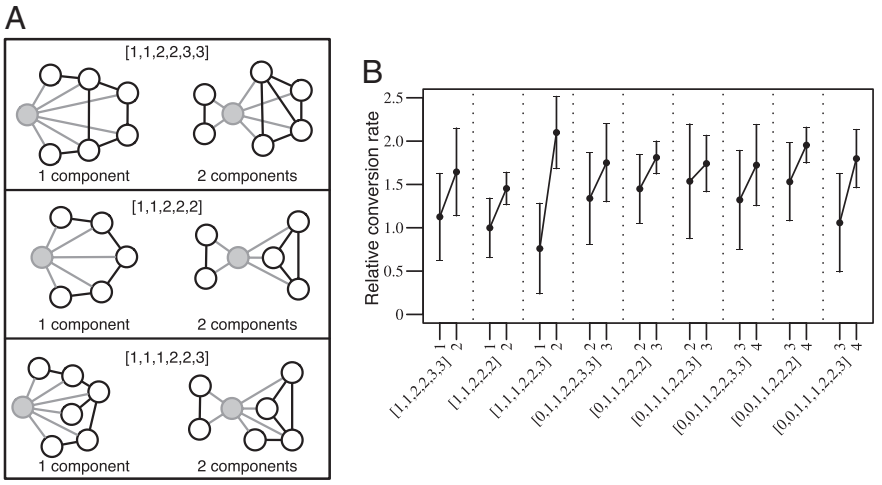

Fig. S3. Recruitment conversion rates for the nine most frequent pairs of graphs with matched embeddedness distributions. (*A*) Illustrations of the three most frequent neighborhood graph pairs with identical embeddedness distributions but differing component counts. The invited node is shaded, and the embeddedness distribution refers to the embeddedness of the shaded edges. The other six frequent graph topology pairs all contain one of these pairs, up to the addition of a node singleton. (*B*) The importance of diversity when controlling for embeddedness, examining the nine most common neighborhood graph pairs with identical embeddedness distributions but differing component counts. Each data point is labeled by its degree distribution and its connected component count, as in Fig. S1. The conversion scale is the same as for the recruitment figures in the main text. Error bars are 95% confidence intervals.

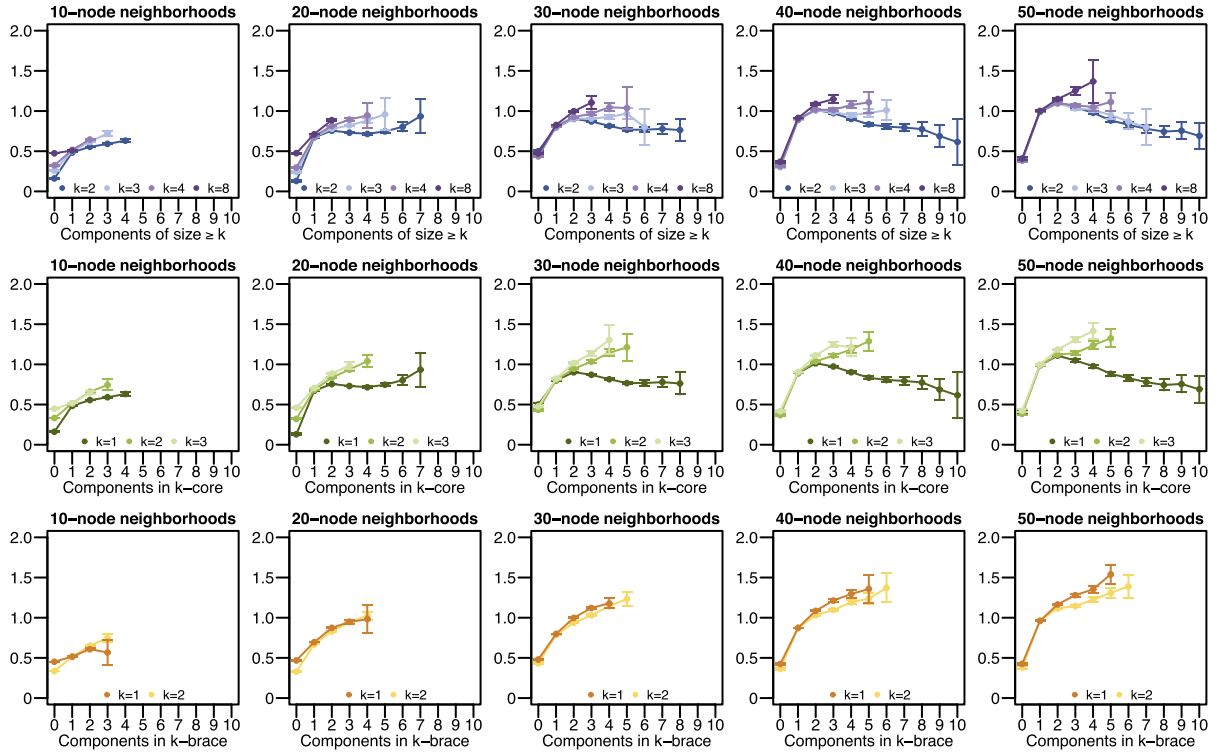

Fig. S4. Engagement as a function of diversity in a neighborhood, conditioned on size. For each size $n = 10, 20, 30, 40, 50$, plots are shown that correspond to Fig. 4 *D–F* in the main text, showing the relative engagement rate as a function of component counts. The 50-node neighborhood plots correspond exactly to the plots in Fig. 4 *D–F*. All engagement rates are reported on a single relative scale, where 1.0 signifies the average conversion rate across all 50-node neighborhoods. Error bars are 95% confidence intervals.

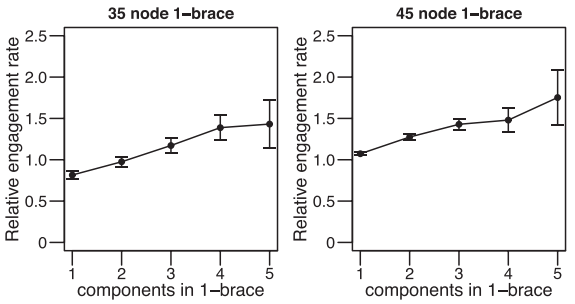

Fig. S5. Controlling for the size of the *k*-brace. We focus on neighborhoods of size 50 with exactly 35 and 45 nodes in their 1-brace and again see that engagement is an increasing function of 1-brace component count. All engagement rates are reported on a single relative scale, where 1.0 signifies the average conversion rate across all 50-node neighborhoods. Error bars are 95% confidence intervals.