# Structural diversity in social contagion

Johan Ugander[a], Lars Backstrom[b], Cameron Marlow[b], and Jon Kleinberg[c,1]

[a]Center for Applied Mathematics and [c]Department of Computer Science, Cornell University, Ithaca, NY 14853; and [b]Facebook, Menlo Park, CA 94025

Edited by Ronald L. Graham, University of California at San Diego, La Jolla, CA, and approved February 21, 2012 (received for review October 6, 2011)

**The concept of contagion has steadily expanded from its original grounding in epidemic disease to describe a vast array of processes that spread across networks, notably social phenomena such as fads, political opinions, the adoption of new technologies, and financial decisions. Traditional models of social contagion have been based on physical analogies with biological contagion, in which the probability that an individual is affected by the contagion grows monotonically with the size of his or her "contact neighborhood"—the number of affected individuals with whom he or she is in contact. Whereas this contact neighborhood hypothesis has formed the underpinning of essentially all current models, it has been challenging to evaluate it due to the difficulty in obtaining detailed data on individual network neighborhoods during the course of a large-scale contagion process. Here we study this question by analyzing the growth of Facebook, a rare example of a social process with genuinely global adoption. We find that the probability of contagion is tightly controlled by the number of connected components in an individual's contact neighborhood, rather than by the actual size of the neighborhood. Surprisingly, once this "structural diversity" is controlled for, the size of the contact neighborhood is in fact generally a negative predictor of contagion. More broadly, our analysis shows how data at the size and resolution of the Facebook network make possible the identification of subtle structural signals that go undetected at smaller scales yet hold pivotal predictive roles for the outcomes of social processes.**

social networks | systems

Social networks play host to a wide range of important social and nonsocial contagion processes (1–8). The microfoundations of social contagion can, however, be significantly more complex, as social decisions can depend much more subtly on social network structure (9–17). In this study we show how the details of the network neighborhood structure can play a significant role in empirically predicting the decisions of individuals.

We perform our analysis on two social contagion processes that take place on the social networking site Facebook: the process whereby users join the site in response to an invitation e-mail from an existing Facebook user (henceforth termed "recruitment") and the process whereby users eventually become engaged users after joining (henceforth termed "engagement"). Although the two processes we study formally pertain to Facebook, their details differ considerably; the consistency of our results across these differing processes, as well as across different national populations (*Materials and Methods*), suggests that the phenomena we observe are not specific to any one modality or locale.

The social network neighborhoods of individuals commonly consist of several significant and well-separated clusters, reflecting distinct social contexts within an individual's life or life history (18–20). We find that this multiplicity of social contexts, which we term structural diversity, plays a key role in predicting the decisions of individuals that underlie the social contagion processes we study.

We develop means of quantifying such structural diversity for network neighborhoods, broadly applicable at many different scales. The recruitment process we study primarily features small neighborhoods, but the on-site neighborhoods that we study in the context of engagement can be considerably larger. For small neighborhoods, structural diversity is succinctly measured by the number of connected components of the neighborhood. For larger neighborhoods, however, merely counting connected components fails to distinguish how substantial the components are in their size and connectivity. To determine whether the structural diversity of on-site neighborhoods is a strong predictor of on-site engagement, we evaluate several variations of the connected component concept that identify and enumerate substantial structural contexts within large neighborhood graphs. We find that all of the different structural diversity measures we consider robustly predict engagement. For both recruitment and engagement, structural diversity emerges as an important predictor for the study of social contagion processes.

## Results

**User Recruitment.** To study the spread of Facebook as it recruits new members, we require information not just about Facebook's users but also about individuals who are not yet users. Thus, suppose that an individual $A$ is not a user of Facebook; it is still possible to identify a set of Facebook users that $A$ may know because these users have all imported $A$'s e-mail address into Facebook. We define this set of Facebook users possessing $A$'s e-mail address to be $A$'s contact neighborhood in Facebook. This contact neighborhood is the subset of potential future friendship ties that can be determined from the presence of $A$'s e-mail address (Fig. 1*A*). Whereas $A$ may in fact know many other people on Facebook as well, such additional friendship ties remain unknown for individuals who do not choose to register and so cannot be studied as a predictor of recruitment. The e-mail contact neighborhoods we study are generally quite small, typically on the order of five or fewer nodes.

We can now study an individual's decision to join Facebook as follows. Facebook provides a tool through which its users can e-mail friends not on Facebook to invite them to join; such an e-mail invitation contains not only a presentation of Facebook and a profile of the inviter, but also a list of the other members of the individual's contact neighborhood. We analyze a corpus of 54 million such invitation e-mails, and the fundamental question we consider is the following: How does an individual's probability of accepting an invitation depend on the structure of his or her contact neighborhood?

Traditional hypotheses suggest that this probability should grow monotonically in the size of the contact neighborhood (3, 9, 10). What we find instead, however, is a striking stratification of acceptance probabilities by the number of connected components in the contact neighborhood (Fig. 1 *B–D* and Fig. S1). When going beyond component count, one may suspect that edge density has a significant impact on the recruitment conversion rate: Among the single-component neighborhoods of a given size, there is a considerable structural difference between neighborhoods connected as a tree and those connected as a clique. However, within the controlled conditional datasets of

Author contributions: J.U., L.B., C.M., and J.K. designed research; J.U., L.B., C.M., and J.K. performed research; J.U., L.B., C.M., and J.K. contributed new reagents/analytic tools; J.U., L.B., C.M., and J.K. analyzed data; and J.U. and J.K. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: kleinber@cs.cornell.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1116502109/-/DCSupplemental.

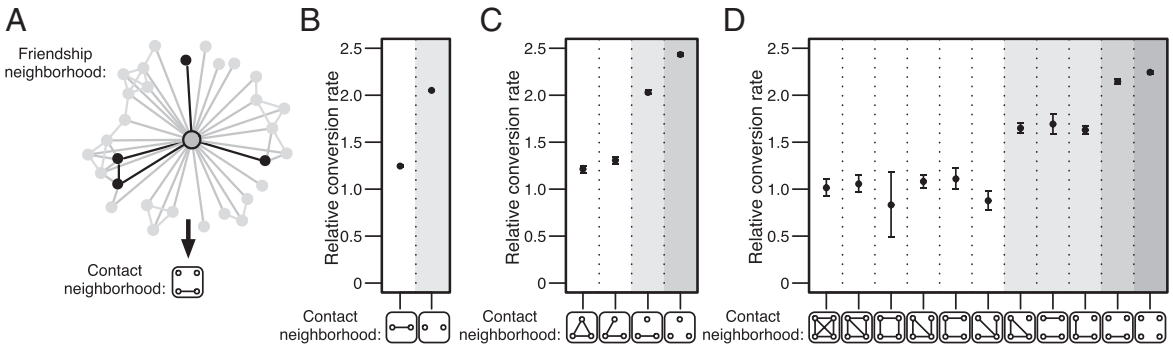


**Fig. 1.** Contact neighborhoods during recruitment. (*A*) An illustration of a small friendship neighborhood and a highlighted contact neighborhood consisting of four nodes and three components. (*B–D*) The relative conversion rates for two-node, three-node, and four-node contact neighborhood graphs. Shading indicates differences in component count. For five-node neighborhoods, see Fig. S1. Invitation conversion rates are reported on a relative scale, where 1.0 signifies the conversion rate of one-node neighborhoods. Error bars represent 95% confidence intervals and implicitly reveal the relative frequency of the different topologies.

one-component neighborhoods of sizes 4–6, we see that edge density has no discernible effect (Fig. 2*A*).

Moreover, we see that once component count is controlled for (Fig. 2*B*), neighborhood size is largely a negative indicator of conversion. In effect, it is not the number of people who have invited you, nor the number of links among them, but instead the number of connected components they form that captures your probability of accepting the invitation. Note that this analysis has been performed in aggregate and thus unavoidably reflects the decisions of different individuals. The ability to reliably estimate acceptance probabilities as a function of something as specific as the precise topology of the contact neighborhood is possible only because the scale of the dataset provides us with sufficiently many instances of each possible contact neighborhood topology (up through size 5).

We view the component count as a measure of "structural diversity," because each connected component of an individual's contact neighborhood hints at a potentially distinct social context in that individual's life. Under this view, it is the number of distinct social contexts represented on Facebook that predicts the probability of joining. We show that the effect of this structural diversity persists even when other factors are controlled for. In particular, the number of connected components in the contact neighborhood remains a predictor of invitation acceptance even when restricted to individuals whose neighborhoods are demographically homogeneous (in terms of sex, age, and nationality; Fig. S2), thus controlling for a type of demographic diversity that is potentially distinct from structural diversity. The component count also remains a predictor of acceptance even when we compare neighborhoods that exhibit precisely the same mixture of "bridging" and "embedded" links (Fig. S3), the key distinction in sociological arguments based on information novelty (19, 20).

For contact neighborhoods consisting of two nodes, we observe that the probability an invitation is accepted is much higher when the two nodes in the neighborhood are not connected by a link (hence forming two connected components, Fig. 1*B*) compared with when they are connected (forming one component). Is there a way to identify cases where people are likely to know each other, even if they are not linked on Facebook? The photo tagging feature on Facebook suggests such a mechanism. Photographs uploaded to Facebook are commonly annotated by users with "tags" denoting the people present in the photographs. We can use these tags to deduce whether two unlinked nodes in a contact neighborhood have been jointly tagged in any photos, a property we refer to as "co-tagging," which serves as an indication of a social tie through copresence at an event (21).

Using photo co-tagging, we find strong effects even in cases where the presence of a friendship tie is only implicit. If a contact neighborhood consists of two unlinked nodes that have nevertheless been co-tagged in a photo, then the invitation acceptance probability drops to approximately what it is for a neighborhood of two linked nodes (Fig. 2*C*). In other words, being co-tagged in a photo indicates roughly the same lack of diversity as being connected by a friendship link. We interpret this result as further evidence that diverse endorsement is key to predicting recruitment. Meanwhile, when the two nodes are friends, co-tags offer a proxy for tie strength, and we see that if the two nodes have also been co-tagged, then the probability of an accepted invitation decreases further. From this we can interpret tie strength as an

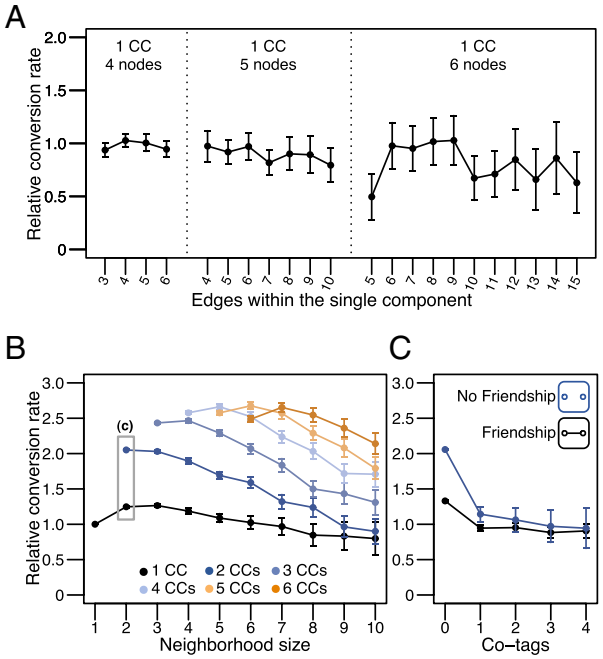


**Fig. 2.** Recruitment contact neighborhoods and component structure. (*A*) Conversion as a function of edge count neighborhoods with one connected component (1 CC) with four to six nodes, where variations in edge count predict no meaningful difference in conversion. (*B*) Conversion as a function of neighborhood size, separated by CC count. When component count is controlled for, size is a negative indicator of conversion. (*C*) Conversion as a function of tie strength in two-node neighborhoods, measured by photo co-tags, a negative indicator of predicted conversion. Recruitment conversion rates are reported on a relative scale, where 1.0 signifies the conversion rate of one-node neighborhoods. Error bars represent 95% confidence intervals.

extension of context, because two strongly tied nodes plausibly constitute an even less diverse endorsement neighborhood.

Finally, we study the position of the inviter within the neighborhood topologies. When studying recruitment, one might suspect that the structural position of the inviter—the person who extended the invitation—might signify differences in tie strength with the invitee and therefore might significantly affect the predicted conversion rate. We find that inviter position figures only slightly in the conversion rate (Fig. 3), with invitations stemming from a high-degree position in the contact neighborhood predicting only a slightly higher conversion rate than if the inviter is a peripheral node.

**User Engagement.** Participation in a social system such as Facebook is built upon a spectrum of social decisions, beginning with the decision to join (recruitment) and continuing on to decisions about how to choose a level of engagement. We now show how structural diversity also plays an analogous role in this latter type of decision process, studying long-term user engagement in the Facebook service. Whereas recruitment is a function of the complex interplay between multiple acts of endorsement, engagement is a function of the social utility a user derives from the service. Our study of engagement focuses on users who registered for Facebook during 2010, analyzing the diversity of their social neighborhoods 1 week after registration as a basis for predicting whether they will become highly engaged users 3 months later.

Users are considered engaged at a given time point if they have interacted with the service during at least 6 of the last 7 days. Facebook had 845 million monthly active users on December 31, 2011, and during the month of December 2011, an average of 360 million users were active on at least 6 out of the last 7 days. We define engagement on a weekly timescale to stabilize the considerable weekly variability of user visits. Our goal is therefore to predict whether a newly registered user will visit Facebook at least 6 of 7 days per week 3 months after registration.

Friendship neighborhoods on Facebook are significantly larger than the e-mail contact neighborhoods from our recruitment study. We focus our engagement study on a population of ∼10 million users who registered during 2010 and had assembled neighborhoods consisting of exactly 10, 20, 30, 40, or 50 friends 1 week after registration. For social network neighborhoods of this size, we find that a neighborhood containing a large number of connected components primarily indicates a large number of one-node components, or "singletons", and as such, it is not an accurate reflection of social context diversity.

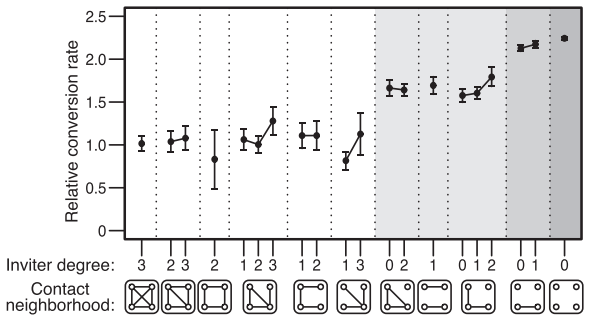


**Fig. 3.** Inviter position during recruitment. Shown is recruitment conversion as a function of neighborhood graph topology and inviter position in neighborhoods of size 4. The position of the inviter within the neighborhood graph is described exactly (up to symmetries) by node degree. Shading indicates differences in component count. Recruitment conversion rates are reported on a relative scale, where 1.0 signifies the conversion rate of one-node neighborhoods. Error bars represent 95% confidence intervals.

To address this, we evaluate three distinct parametric generalizations of component count. First, we measure diversity simply by considering only components over a certain size $k$. Second, we measure diversity by the component count of the *k-core* of the neighborhood graph (22), the subgraph formed by repeatedly deleting all vertices of degree less than $k$. Third, we define a measure that isolates dense social contexts by removing edges according to their *embeddedness*, the number of common neighbors shared by their two endpoints; intuitively this is an analog, for edges, of the type of node removal that defines the $k$-core. Adapting earlier work on embeddedness by Cohen (23), we define the *k-brace* of a graph to be the subgraph formed by repeatedly deleting all edges of embeddedness less than $k$ and then deleting all single-node connected components. (Cohen's work was concerned with a definition equivalent to the largest connected component of the $k$-brace; because we deal with the full subgraph of all nontrivial components, it is useful to adapt the definitions as needed.) Examples of these three measures applied to a neighborhood graph are shown in Fig. 4 *A* and *B*, illustrating the

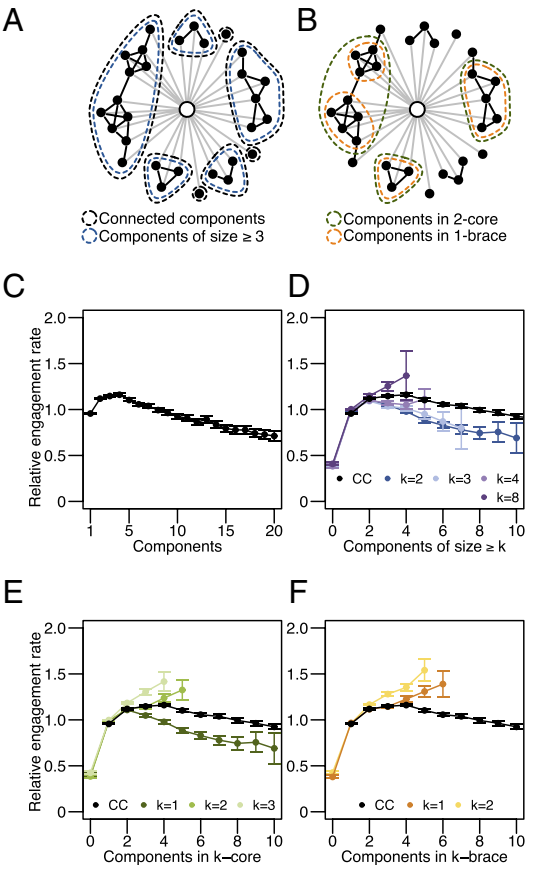


**Fig. 4.** Engagement and structural diversity for 50-node friendship neighborhoods. (*A*) Illustration of the connected components in a friendship neighborhood, delineating connected components and components of size ≥3. (*B*) Illustration of the $k$-core and the $k$-brace, delineating the connected components of the 2-core and the 1-brace. (*C*) Engagement as a function of connected component count. (*D*) Engagement as a function of the number of components of size ≥ $k$, for $k = 2, 3, 4, 8$, with connected component (CC) count shown for comparison. (*E*) Engagement as a function of $k$-core component count for $k = 1, 2, 3$, with CC count shown for comparison. (*F*) Engagement as a function of $k$-brace component count for $k = 1, 2$, with CC count shown for comparison. Engagement rates are reported on a relative scale, where 1.0 signifies the average conversion rate of all 50-node neighborhoods. All error bars are 95% confidence intervals. For other neighborhood sizes, see Fig. S4.

connected components of size 3 or greater, the connected components of the 2-core, and the connected components of the 1-brace. We see that the three parametric measures we evaluate differ measurably in how they isolate "substantial" social contexts.

The $k$-core component count for $k = 0$ is simply the component count of the original graph, the same as we analyzed when examining recruitment. For $k = 1$, the $k$-core component count is the count of nonsingleton components, whereas for $k = 2$, all tree-like components are discarded and the remaining components are counted. When considering the $k$-brace, observe that for all graphs the $k$-brace is a subgraph of the $(k + 1)$-core: indeed, because each node in the $k$-brace is incident to at least one edge, and each edge in the $k$-brace has embeddedness at least $k$, all nodes in the $k$-brace must have degree at least $k + 1$. It is therefore reasonable to compare the 1-brace to the 2-core. Both of these restrictions discard tree-like components, but the 1-brace will tend to break up components further than the 2-core does—the operation defining the 1-brace continues to cleave components in cases where sets of nodes forming triangles are linked together by unembedded edges or where a component contains cycles but no triangles. The notion of the $k$-core has been applied both to the study of critical phenomena in random graphs (24, 25) and to models of the Internet (26, 27), but to our knowledge the $k$-brace has not been studied extensively (see *SI Text* for some basic results on the $k$-brace and ref. 23 for analysis of a related definition).

When studying the structural diversity of 1-week Facebook friendship neighborhoods as a predictor of long-term engagement, simply counting connected components leads to a muddled view of predicted engagement (Fig. 4*C*). However, extending the notion of diversity according to any of the definitions above suffices to provide positive predictors of future long-term engagement. Specifically, when considering the components of the 1-brace, which removes small components and severs unembedded edges, we see that diversity (captured by the presence of multiple components) emerges as a significant positive predictor of future long-term engagement (Fig. 4*F*). We also see that the closely related 2-core component count is a clean predictor (Fig. 4*E*). Finally, if we consider simply the number of components of size $k$ or larger in the original neighborhood (without applying the core or brace definitions), we see that small values of $k$ are not enough (Fig. 4*D*); but even here, when $k$ is increased to make the selection over components sufficiently astringent (in particular, when we count only components of size 8 or larger), a clean indicator of engagement again emerges.

When considering the $k$-brace, it is sufficient to consider the component count of the 1-brace for our purposes, but larger values of $k$ may be useful for analyzing larger neighborhoods in other domains. We note that the presence of several components in the $k$-core and the $k$-brace is fundamentally limited by the size of the core/brace, and we perform a control of this potentially confounding factor (Fig. S5). The conventional wisdom for social systems such as Facebook is that their utility depends crucially upon the presence of a strong social context. Our findings validate this view, observing that the predicted engagement for users who lack any strong context (e.g., those who have zero components in their neighborhood 1-brace) is much lower than for those with such a context. Our analysis importantly extends this view, finding that the presence of multiple contexts introduces a sizable additional increase in predicted engagement.

A cruder approach to diversity might consider measuring diversity through the edge density of a neighborhood, figuring that sparse neighborhoods would be more varied in context. In Fig. 5 we see how this approach results in a complicated view where the optimal edge density for predicting engagement lies at an internal and size-dependent optimum. Given what our component analysis reveals, we interpret this observation as a superposition of two effects: Too few edges imply a lack of context (4) but too many edges imply a lacking diversity of contexts, with a nontrivial interior clearly dominating the boundary conditions. From Fig. 5 it also becomes clear that internal neighborhood structure is at least as important as size, with a 20-node neighborhood featuring a well-balanced density predicting higher conversion than a sparse or dense 50-node neighborhood.

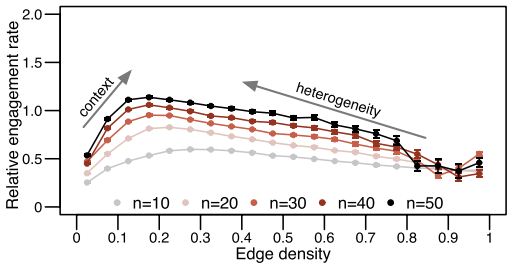


**Fig. 5.** Engagement as a function of edge density. For five different neighborhood sizes, $n = 10, 20, 30, 40, 50$, we see that when component count is not accounted for, an internal engagement optimum is observed, showing the combined forces of focused context and structural heterogeneity. Engagement rates are reported on a relative scale, where 1.0 signifies the average conversion rate of all 50-node neighborhoods. All error bars are 95% confidence intervals.

## Discussion

Detailed traces of Facebook adoption provide natural sources of data for studying social contagion processes. Our analysis provides a high-resolution view of a massive social contagion process as it unfolded over time and suggests a rethinking of the underlying mechanics by which such processes operate. Rather than treating a person's number of neighbors as the crucial parameter, consider instead the number of distinct social contexts that these neighbors represent as the driving mechanism of social contagion.

The role of neighborhood diversity in contagion processes suggests interesting further directions to pursue, both for mathematical modeling and for potential broader applications. Mathematical models in areas including interacting particle systems (28, 29) and threshold contagion (3, 30) have explored some of the global phenomena that arise from contagion processes in networks for which the behavior at a given node has a nontrivial dependence on the full set of behaviors at neighboring nodes. Neighborhood diversity could be naturally incorporated into such models by basing the underlying contagion probability, for example, on the number of connected components formed by a node's affected neighbors. It then becomes a basic question to understand how the global properties of these processes change when such factors are incorporated.

More broadly, across a range of further domains, these findings suggest an alternate perspective for recruitment to political causes, the promotion of health practices, and marketing; to convince individuals to change their behavior, it may be less important that they receive many endorsements than that they receive the message from multiple directions. In this way, our findings propose a potential revision of core theories for the roles that networks play across social and economic domains.

## Materials and Methods

**Recruitment Data Collection.** Here we discuss details of the e-mail recruitment data. All user data were analyzed in an anonymous, aggregated form. The contact neighborhood individuals included in invitation e-mails are limited to nine in number, and so we have restricted our analysis to neighborhoods (inviter plus contact importers) of 10 nodes or less. In cases with more than nine candidate "other people you may know," the invitation tool selects a randomized subset of nine for inclusion in the e-mail.

We conditioned our data collection upon several criteria. First, we considered only first invitations to join the site. Subsequent invitations to an e-mail address are handled differently by the invitation tool, and so we have not included them in our study. Second, we considered only invitations where

the inviter invited at most 20 e-mail addresses on the date of the invitation. This conditioning is meant to omit invitation batches where the inviter opted to "select all" within the contact import tool and focuses our investigation on socially selective invitations.

Invitations were sent during an 11-week period spanning July 12, 2010 to September 26, 2010. An e-mail address was considered to have converted to a registered user account if the address was registered for an account within 14 days of the invitation, counting both individuals who signed up via links provided in the invitation e-mail and users who signed up by visiting the Facebook website directly within 14 days. Only contact import events that occurred before the invitation event are considered. Likewise, only friendship edges that existed before the invitation event are considered to be part of the neighborhood.

Many of the findings we investigate are governed by complex nonlinear effects, which make traditional regression controls generally inadequate. In an attempt to control for confounding signals in our data, several parallel observation groups were maintained, against which all findings were validated. As a means of capturing potential artifacts from duplicitous private/business e-mail address use, a first such validation group was constructed by conditioning upon e-mail invitations sent to a small set of common and commonly private e-mail providers: Hotmail, Yahoo!, Gmail, AOL, and Yahoo! France. As a means of observing any differences between already established and growing Facebook markets, two parallel validation groups were constructed to observe established markets (United States) and emerging Facebook markets (Brazil, Germany, Japan, and Russia), classified by the most recently resolved country of login for the inviting Facebook account. Whereas invitation conversion rates were generally higher in emerging markets, none of the conditional datasets were observed to deviate from the complete dataset with regard to internal structural findings.

Highly sparse neighborhoods were a very common occurrence in these data, owing to the fact that the neighborhoods we study here are only partial observations of an individual's actual connection to Facebook. We are able to infer links only to those site users who have used the contact importer tool and maintain active e-mail communication with the e-mail address in question, criteria that induce a sampled subgraph that we then observe. The probability of sampling an edge uniformly at random in any neighborhood with low edge density is therefore quite low, and the probability that all sampled nodes come from the same cluster within a clustered neighborhood is lower still. From the perspective of communication multiplexity (31), we should in fact expect that our randomly induced subgraph sample is biased toward strongly connected ties that tend to communicate on multiple mediums, but this expectation is not at issue with our results. The real matter of the fact is that contact neighborhoods where the induced subgraph consists of a single connected component are likely to come from *very* tightly connected neighborhood graphs.

Although the contact importer tool and invitation tool are prominently featured as part of the new user experience on Facebook, they are also heavily used by experienced users of the site: The median site age of an inviter in our dataset was 262 days. Although e-mail invitations constitute only a small portion of Facebook's growth, they provide a valuable window into the otherwise invisible growth process of the Facebook product.

For the analysis of photo co-tags, only co-tags since January 1, 2010 were considered.

**Engagement Data Collection.** We consider users *engaged* at a given time point if they have interacted with the application during at least 6 of the last 7 days. As with any measure of user behavior, this metric is a heuristic merely meant to approximate a broader notion of involvement on the site. Highly engaged users who do not access the Internet on weekends will never qualify as "six-plus engaged," whereas users who simply log in on a daily basis to check their messages will qualify. Our analysis is restricted to the population level, so such confounders are not a problem.

Due to the technical nature of how engagement data are stored at Facebook, it is impractical to retrieve six-plus engagement measures for dates exactly 3 months after registration. As an appropriate surrogate, we consider the six-plus engagement of users on the first day of their third calendar month as users.

**ACKNOWLEDGMENTS.** We thank M. Macy, J. Fowler, D. Watts, and S. Strogatz for comments. This research has been supported in part by a MacArthur Foundation Fellowship and National Science Foundation Grants IIS-0705774, IIS-0910664, CCF-0910940, and IIS-1016099.

1. Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. *Phys Rev Lett* 86:3200–3203.
2. Newman ME, Watts DJ, Strogatz SH (2002) Random graph models of social networks. *Proc Natl Acad Sci USA* 99(Suppl 1):2566–2572.
3. Dodds PS, Watts DJ (2004) Universal behavior in a generalized model of contagion. *Phys Rev Lett* 92:218701.
4. Backstrom L, Huttenlocher D, Kleinberg J, Lan X (2006) Group formation in large social networks: Membership, growth, and evolution. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, eds Eliassi-Rad T, Ungar LH, Craven M, Gunopulos D (Association for Computing Machinery, New York), pp 44–54.
5. Kearns M, Suri S, Montfort N (2006) An experimental study of the coloring problem on human subject networks. *Science* 313:824–827.
6. Watts DJ, Dodds PS (2007) Influentials, networks, and public opinion formation. *J Consum Res* 34:441–458.
7. Christakis NA, Fowler JH (2007) The spread of obesity in a large social network over 32 years. *N Engl J Med* 357:370–379.
8. Sun E, Rosenn I, Marlow C, Lento T (2009) Gesundheit! Modeling contagion through Facebook news feed. *Proceedings of the AAAI International Conference on Weblogs and Social Media*, eds Adar E, et al. (Association for the Advancement of Artificial Intelligence, Menlo Park, CA), pp 146–153.
9. Schelling T (1971) Dynamic models of segregation. *J Math Sociol* 1:143–186.
10. Granovetter M (1978) Threshold models of collective action. *Am J Sociol* 83: 1420–1443.
11. Burt R (1987) Social contagion and innovation: Cohesion versus structural equivalence. *Am J Sociol* 92:1287–1335.
12. Kossinets G, Watts DJ (2006) Empirical analysis of an evolving social network. *Science* 311:88–90.
13. Centola D, Eguíluz V, Macy M (2007) Cascade dynamics of complex propagation. *Physica A* 374:449–456.
14. Centola D, Macy M (2007) Complex contagions and the weakness of long ties. *Am J Sociol* 113:702–734.
15. Palla G, Barabási AL, Vicsek T (2007) Quantifying social group evolution. *Nature* 446: 664–667.
16. Aral S, Muchnik L, Sundararajan A (2009) Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc Natl Acad Sci USA* 106: 21544–21549.
17. Fowler JH, Christakis NA (2010) Cooperative behavior cascades in human social networks. *Proc Natl Acad Sci USA* 107:5334–5338.
18. Simmel G (1955) *Conflict and the Web of Group Affiliations*, eds trans Wolff K, Bendix R (Free Press, Glencoe, IL).
19. Granovetter M (1973) The strength of weak ties. *Am J Sociol* 78:1360–1380.
20. Burt R (1992) *Structural Holes: The SocialStructure of Competition* (Harvard Univ Press, Cambridge, MA).
21. Crandall D, et al. (2010) Inferring social ties from geographic coincidences. *Proc Natl Acad Sci USA* 107:22436–22441.
22. Bollobás B (2001) *Random Graphs* (Cambridge Univ Press, Cambridge, UK), 2nd Ed, p 150.
23. Cohen JD (2008) Trusses: Cohesive subgraphs for social network analysis. *National Security Agency Technical Report* (National Security Agency, Fort Meade, MD).
24. Luczak T (1991) Size and connectivity of the k-core of a random graph. *Discrete Math* 91:61–68.
25. Janson S, Luczak MJ (2007) A simple solution to the k-core problem. *Random Struct Algo* 30:50–62.
26. Alvarez-Hamelin JI, Dall'Astra L, Barrat A, Vespignani A (2006) Large scale networks fingerprinting and visualization using the k-core decomposition. *Adv Neural Inf Process Syst* 18:41–50.
27. Carmi S, Havlin S, Kirkpatrick S, Shavitt Y, Shir E (2007) A model of Internet topology using k-shell decomposition. *Proc Natl Acad Sci USA* 104:11150–11154.
28. Liggett T (1985) *Interacting Particle Systems* (Springer, Berlin).
29. Durrett R (1995) *Ten Lectures on Particle Systems* (Springer, Berlin).
30. Mossel E, Roch S (2007) On the submodularity of influence in social networks. *Proceedings of the ACM Symposium on Theory of Computing*, eds Johnson DS, Feige U (Association for Computing Machinery, New York), pp 128–134.
31. Haythornthwaite C, Wellman B (1998) Work, friendship, and media use for information exchange in a networked organization. *J Am Soc Inf Sci* 49:1101–1114.

# 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).

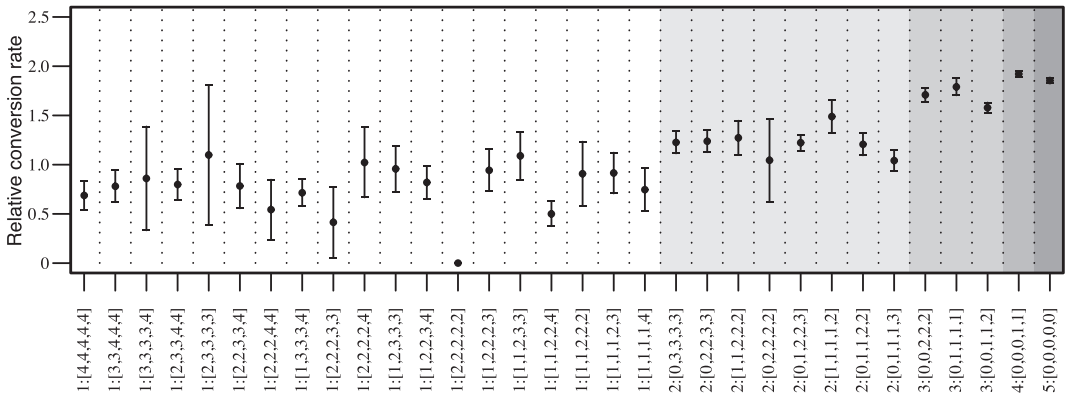


**Fig. S1.** Invitation conversion rates of size five neighborhoods, labeled by their number of connected components and their degree distribution, which uniquely determines the topology (for graphs of size 4 or smaller, degree distribution is a unique determinant of component count, but this uniqueness 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.

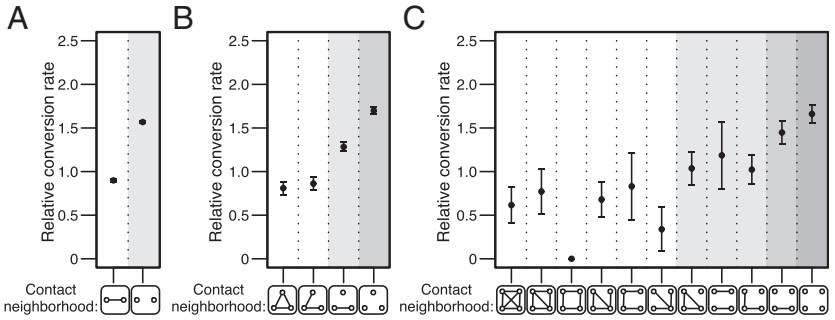


**Fig. S2.** Recruitment conversion for demographically homogeneous neighborhoods, as a function of (*A*) two-node, (*B*) three-node, and (*C*) four-node contact neighborhood graphs. The conversion scale is the same as for Fig. 1 in the main text. Error bars represent 95% confidence intervals.

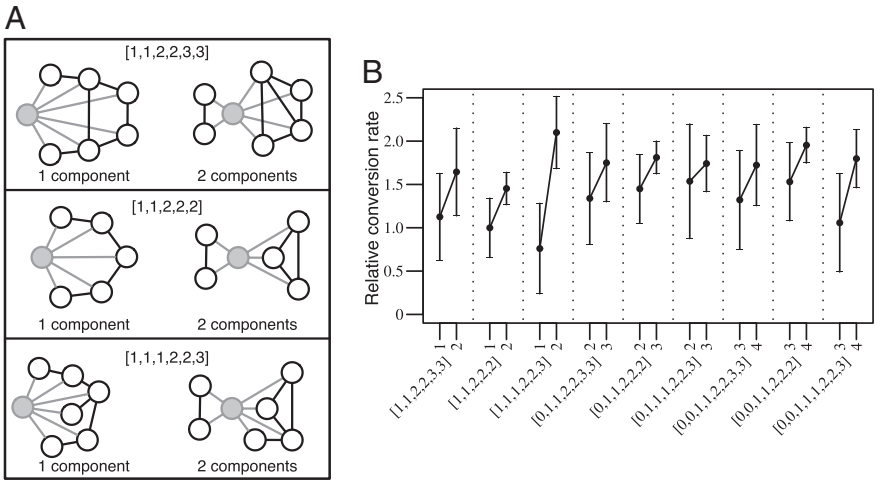


**Fig. S3.** Recruitment conversion rates for the nine most frequent pairs of graphs with matched embeddedness distributions. (*A*) Illustrations of the three most frequent neighborhood graph pairs with identical embeddedness distributions but differing component counts. The invited node is shaded, and 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.

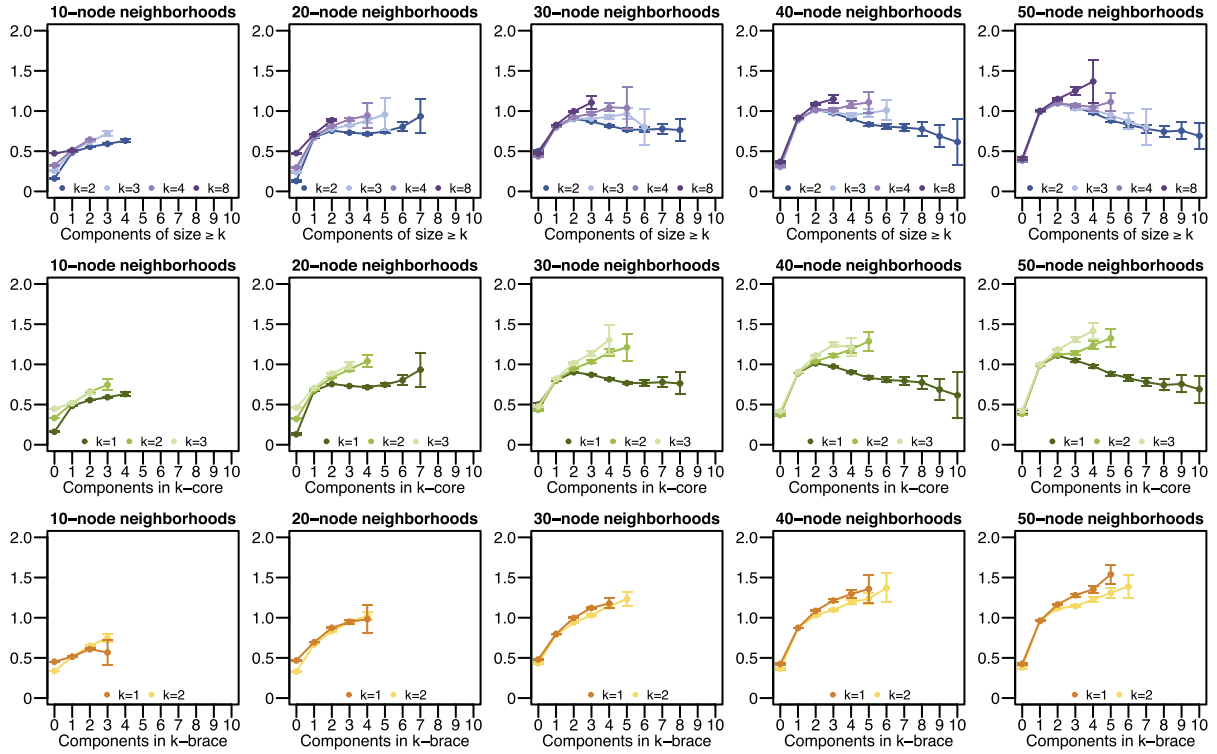


**Fig. S4.** Engagement as a function of diversity in a neighborhood, conditioned on size. For each size $n = 10, 20, 30, 40, 50$, plots are shown that correspond to Fig. 4 *D–F* in the main text, showing the relative engagement rate as a function of component counts. The 50-node neighborhood plots correspond 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.

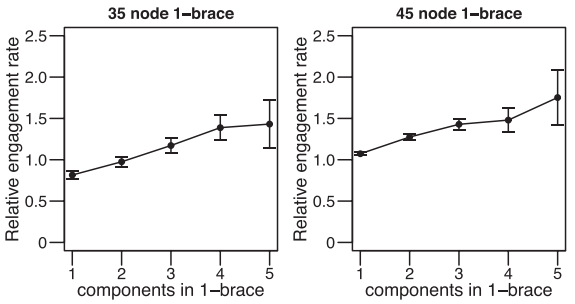


**Fig. S5.** Controlling for the size of the *k*-brace. We focus on neighborhoods of size 50 with exactly 35 and 45 nodes in their 1-brace and again see that engagement is an increasing function of 1-brace component count. All engagement rates are reported on a single relative scale, where 1.0 signifies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