nature human behaviour

LETTERS

https://doi.org/10.1038/s41562-018-0346-z

Corrected: Publisher Correction

Social influence maximization under empirical influence models

Sinan Aral* and Paramveer S. Dhillon*

Social influence maximization models aim to identify the smallest number of influential individuals (seed nodes) that can maximize the diffusion of information or behaviours through a social network. However, while empirical experimental evidence has shown that network assortativity and the joint distribution of influence and susceptibility are important mechanisms shaping social influence, most current influence maximization models do not incorporate these features. Here, we specify a class of empirically motivated influence models and study their implications for influence maximization in six synthetic and six real social networks of varying sizes and structures. We find that ignoring assortativity and the joint distribution of influence and susceptibility leads traditional models to underestimate influence propagation by 21.7% on average, for a fixed seed set size. The traditional models and the empirical types that we specify here also identify substantially different seed sets, with only 19.8% overlap between them. The optimal seeds chosen under empirical influence models are relatively less well-connected and less central nodes, and they have more cohesive, embedded ties with their contacts. Hence, empirically motivated influence models have the potential to identify more realistic sets of key influencers in a social network and inform intervention designs that disseminate information or change attitudes and behaviours.

Recent debates about the use of empirical data in micro-targeting campaigns designed to change opinions and behaviours in social media networks have focused almost exclusively on the effectiveness of these interventions in changing the opinions and behaviours of the targeted groups. There is, however, a potentially broader application of such empirical data to the optimization of the spread of those behaviours in a social network through influence maximization. In fact, the social 'influence maximization' problem lies at the heart of networks research in multiple disciplines, including physics[1], economics[2], computer science[3,4] and sociology[5]. This elegant problem is essential to our scientific understanding of information diffusion[6], cascade dynamics[7], behavioural contagion[8] and concrete policy decisions in applied fields such as marketing[9,10], contagion management[11], immunization[12,13] and public health[14].

Given a network graph, a seeding budget, an influence model (which governs influence diffusion over the network) and an optimization framework (to select the seed nodes), the goal of social influence maximization is to choose a set of seeds (individuals in the network) to receive an encouragement to adopt a product or behaviour (for example, an advertisement or incentive) such that the 'influence' of the seeds spreads the behaviour to the maximum number of nodes in the network. The problem was first formulated as a probabilistic model of interaction with heuristics for choosing the best seeds[9] and subsequently framed as a discrete optimization problem[3,4]. Although social influence maximization is multifaceted, some dimensions of the problem have attracted more research interest than others.

The optimization framework has received the most attention, as researchers developed efficient discrete optimization strategies for choosing the seed set. The optimization is known to be NP-hard[15] and a greedy algorithm that achieves a $1-1/e$ approximation has been proposed previously[3]. Since then, multiple refinements have improved the computational efficiency of the procedure[16–18] and have implemented optimization in software that substantially reduces the run time of the original greedy algorithm[19–21]. However, the influence model, which specifies the influence diffusion process in the network (that is, how the behaviour of a set of seed nodes at time t diffuses to other nodes at time $t+n$), has received much less attention, except in some recent studies that describe algorithms for robust influence maximization in the presence of uncertainty in edge propagation probabilities or the influence functions[22,23]. Two broad classes of influence models exist in the current literature: threshold models and cascade models.

Threshold models assume that there is a threshold value (or a set of threshold values) at which nodes adopt the product or behaviour. In the simplest of these models, the linear threshold (LT) model[24], each node v has a latent threshold θ_v and for every neighbour $u \in N(v)$ (where $N(v)$ is the set of neighbours in the graph) (u,v) has a non-negative weight w_{uv} such that $\sum_{u \in N(v)} w_{uv} \leq 1$. Given the thresholds and an initial set of active nodes, the process unfolds deterministically in discrete time steps. At time t, an inactive node v becomes active if $\sum_{u \in A(v)} w_{uv} \geq \theta_v$ where $A(v)$ is the set of active neighbours of v up to time-step $t-1$. Once activated, a node stays active and the process terminates when no more activations are possible.

Cascade models were first introduced in the marketing context[25,26]. In the simplest of these models, the independent cascade (IC) model, each edge (u,v) in the graph is associated with a probability p_{uv} known as the influence probability. Given the influence probabilities and an initial set of active nodes, as each node u becomes active, it is given a single chance to activate each inactive neighbour v independently with probability p_{uv}. As in the case of the LT model, the process unfolds in discrete time and if u has multiple newly activated neighbours, their attempts are modelled sequentially in an arbitrary order. The temporal dynamics of the spread of influence can be modelled in discrete or continuous time. Traditionally, the discrete time setting has received the most attention, mainly owing to convenience, although recently some models have proposed continuous time dynamics[27].

These current approaches to influence modelling entrench the view that a node's influence can be characterized either by its network properties or by a transmission parameter that is specified as constant, random or drawn from a uniform distribution. The weight parameters in the LT model (w_{uv}) and the influence

MIT Sloan School of Management, Cambridge, MA, USA. *e-mail: sinan@mit.edu; dhillon@mit.edu

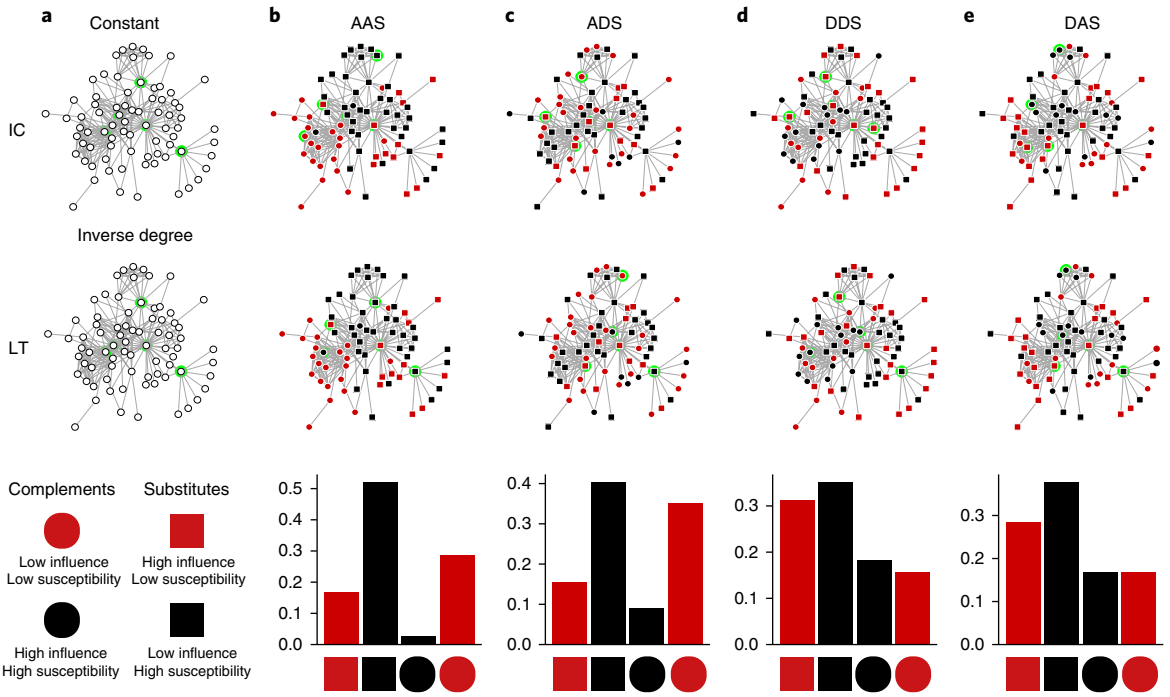

Fig. 1 | Parameterization of influence and susceptibility and implications for seed set selection. The same network is displayed, parameterized by four different models of the distribution of influence and susceptibility over nodes, characterized by four types of nodes: low influence and low susceptibility nodes, high influence and low susceptibility nodes, high influence and high susceptibility nodes and low influence and high susceptibility nodes. The optimal seed nodes selected under each model are outlined in green. **a**, Baseline IC and LT models for which propagation properties are specified as constant (top) and the inverse of node degree (bottom), respectively. **b**, Baseline IC and LT models for which propagation properties are specified according to the assortative influence, assortative susceptibility, substitute influence–susceptibility (AAS) model. **c**–**e**, The same information as in **b**, but for the assortative influence, disassortative susceptibility, substitute influence–susceptibility (ADS; **c**), disassortative influence, disassortative susceptibility, substitute influence–susceptibility (DDS; **d**) and disassortative influence, assortative susceptibility, substitute influence–susceptibility (DAS; **e**) empirical influence models. Distributions of the frequency of the four types of nodes with different influence and susceptibility characterizations are displayed underneath each graph or model. Seed sets differ substantially across different parameterizations of the graph, implying vastly different influence maximization results for the different models of influence and susceptibility.

parameters in the IC model (p_{uv}) are free parameters that need to be fixed. Typically, the w_{uv} parameter in the LT model is chosen as inversely proportional to the in-degree of the node and the p_{uv} parameter in the IC model is assumed to be a constant uniform edge propagation probability of 0.01 or 0.1[3,4]. Although the previous theory[3] is general and simply assumes an unknown parameter p_{uv} between 0 and 1, the experiments in the previous study[3] and in most of the subsequent studies fix p_{uv} to 0.01 or 0.1. Even the most recent work, which maps influence maximization to optimal percolation (a variant of the LT model in which node thresholds are fixed proportional to their degree and all the edge weights are 1), relies on network structure, rather than estimates of the distribution of influence and susceptibility in real networks, to govern diffusion dynamics[28].

Unfortunately, this is not how social influence diffuses in real networks. Empirical evidence suggests that these influence models are misspecified in three important ways. First, real networks are assortative[29]. However, despite evidence showing that assortativity substantially impacts diffusion dynamics in networks[30], even the most recent influence models used for influence maximization (for example, ref. [18]) do not currently model the assortativity of influence or how influence is distributed in the network. Second, most of the recent empirical studies on the identification of peer effects[31], homophily[32] and social influence[33] support modelling of diffusion using a non-uniform joint distribution of influence and susceptibility[34]. Despite this focus on 'influential people' and 'susceptible people' in the empirical literature[35], current influence maximization models do not separate influence from susceptibility or specify their joint distribution. Third, to be realistic, influence models must accommodate heterogeneity in influence and susceptibility. Some LT and IC models incorporate heterogeneity in influence and some do not. For instance, when thresholds (in the LT model) or weights (in the IC model) are distributed uniformly in a random manner, some heterogeneity is incorporated into the influence model specification. However, when these values are specified as constant, influence is assumed to be homogenous; and when the values are specified as proportional to degree, the heterogeneity in influence depends on the heterogeneity of the degree distribution. Incorporating heterogeneity improves current models, but our analysis shows that it is not enough to choose 'optimal' seeds, where optimal is defined under an empirical model of the joint distribution of influence and susceptibility in the network. In the end, all three sources of misspecification (the joint distribution of influence and susceptibility, the assortativity of influence and susceptibility and heterogeneity) have important roles in differentiating empirical influence models from the current models used in influence maximization.

The degree to which influence maximization applies to real policy decisions, such as which customers to market to or which people to immunize, depends almost entirely on whether the influence model is correctly specified. If influence models are realistic and reflect our empirical understanding of influence diffusion, then social influence maximization will produce realistic optimal seed sets that create behavioural diffusion. If influence models are misspecified, however, influence maximization will produce unrealistic and suboptimal seed sets.

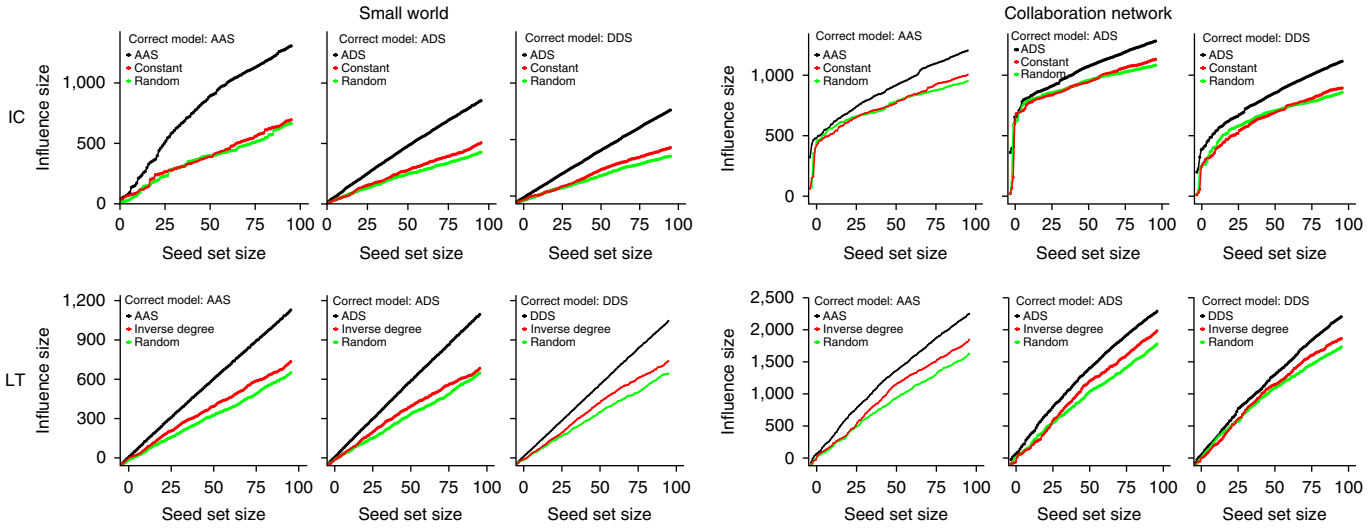

Fig. 2 | Influence diffusion under identical influence maximization regimes with different influence models. The total influence sizes (number of adopter nodes) conditional on the seed set size (the number of initial seed nodes specified by influence maximization) under different influence models (AAS, ADS and DDS) for the IC and LT models in a synthetic small world network (left) and the arXiv high-energy physics collaboration network (right). The total number of adopters is generated by applying influence maximization on the graph given a true underlying influence model (for example, AAS and IC) and then using either that model or the baseline models (inverse degree, random or constant) to maximize influence. The difference in total adopters achieved under different models represents the suboptimality of using a baseline model of influence to choose seed nodes when the true influence model is one of the empirical influence models that we specify based on recent empirical evidence.

Here, we specify a class of empirical influence models and study their implications for social influence maximization in six synthetic and six real social networks of varying sizes and structures. We quantify the extent to which influence model misspecification produces suboptimal seed sets and inaccurate projections of the amount of influence created by optimally chosen policies. The key insights derived from this exercise stem from the two main modelling contributions in our specifications. First, by distinguishing influence and susceptibility, we allow for the possibility that influential people interact with susceptible people as well as with those who are less susceptible to influence. Rather than specifying someone's average influence as a constant transmission parameter that governs their ability to influence everyone they know in the same way, in our specifications someone's influence varies systematically across their contacts and influential people are more effective at spreading influence to susceptible people than to those who are less susceptible. Second, by specifying the joint distribution of influence and susceptibility together, we allow for assortativity in influence and susceptibility in the network and enable investigations of how the distribution of influence and susceptibility over nodes in the network affects diffusion dynamics and influence maximization. In this way, we ask how diffusion and thus the results of influence maximization change when influential people are surrounded by susceptible people, as opposed to, for example, when influential people and susceptible people cluster together, but not with each other. Figure 1 shows the parameterization of our empirical influence models and their impact on seed selection.

The results show that incorporating more realistic diffusion dynamics into the heart of the influence maximization problem leads to vastly different results. In particular, current approaches underestimate influence propagation by 21.7% on average, for a fixed seed set size. Perhaps more importantly, the optimal seed sets under empirical influence models only overlap with optimal seed sets under traditional models by 19.8% on average, indicating that influence maximization procedures under unrealistic influence models rarely select optimal seeds. Moreover, the optimal seeds chosen under empirical influence models are relatively less well-connected (as measured by their degree), are relatively less central nodes and have more cohesive, embedded ties with their contacts, compared to the seeds chosen by baseline methods from the extant influence maximization literature.

These results indicate qualitatively different policy prescriptions from influence maximization. Not only are the optimal seed nodes different under empirical influence models, they are systematically different in ways that enable effective adjustments to influence maximization heuristics as well as to our understanding of what characteristics drive influence maximization in networks.

All variations of the empirical influence model spread substantially more influence than the baseline models (Fig. 2) and seed node selection using models based on the latest empirical evidence substantially outperform seed node selection based on current influence maximization models in all twelve of the graph structures that we studied (representative examples are given in the main manuscript, and the complete set of influence maximization results are presented in the Supplementary InformationInformation). We compare our seed set selections to the inverse degree, random and constant influence model specifications in order to conservatively estimate the suboptimality of current models, as these models represent the current state-of-the-art models of influence maximization. The constant baseline model assigns edge propagation probabilities based on the heuristic of picking a fixed number of 0.1 for all the edges for the IC model. For the LT model, the same heuristic of picking a fixed number reduces to choosing the edge weights based on the inverse of the in-degree of the node, because the incoming weights of each node sum to 1. So, even in the case of the LT model, all the edge weights are constant, however, they are all equal to the inverse of the in-degree of the node.

The results in Fig. 2 show that the correct empirical influence model performs better than the baseline models in all cases and that, in many cases, the suboptimality of the seed sets chosen by baseline models is severe. The fact that the random baseline model performs comparably to the constant and inverse degree baseline models and substantially worse than the correct empirical influence propagation model highlights that the superior performance of the

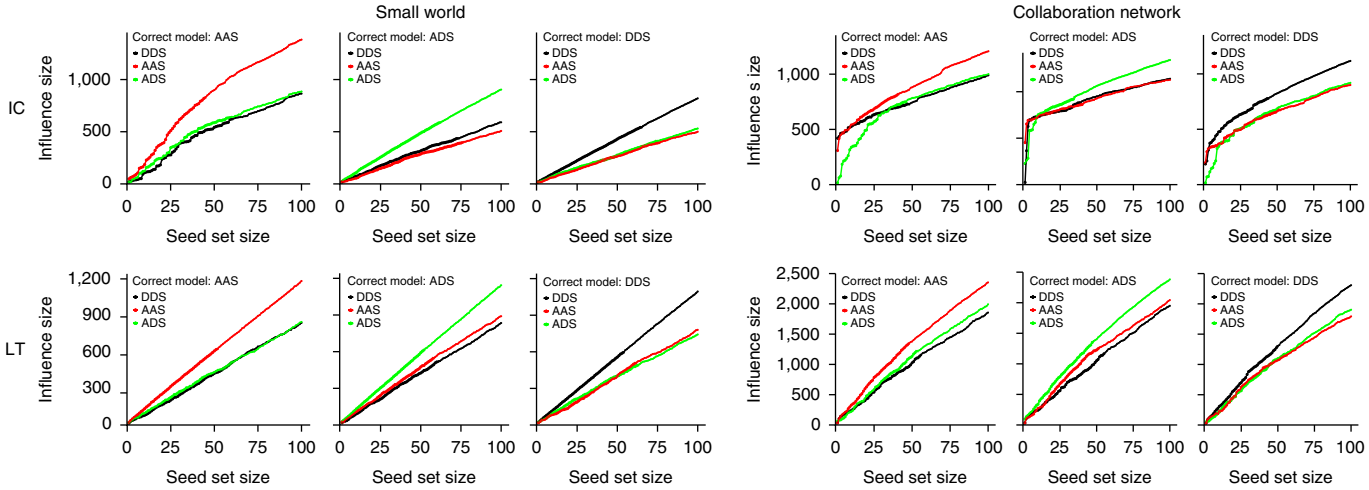

Fig. 3 | Influence diffusion under identical influence maximization regimes with different influence models. The total influence sizes (number of adopter nodes) conditional on the seed set size (the number of initial seed nodes specified by influence maximization) under different influence models (AAS, ADS and DDS) for the IC and LT models in a synthetic small world network (left) and the arXiv high-energy physics collaboration network (right). The total number of adopters is generated by applying influence maximization on the graph given a true underlying influence model (for example, AAS and IC) and then using either that model or one of the other empirical influence models (for example, AAS, ADS or DDS) to maximize influence. The difference in total adopters achieved under different models represents the suboptimality of using a different model of influence to choose seed nodes when the true influence model is the empirical influence model that we specify based on recent empirical evidence. The analysis in this figure is the same as in Fig. 2, with the only difference being that here we compare the true influence model against other incorrect empirical influence models, rather than against the heuristic baselines (inverse degree, random or constant).

correct model is not only because of the randomness it inserts into the edge propagation thresholds, but also because of the imposition of the correct correlation structure of edge propagation thresholds that leads to the selection of higher quality seed nodes and therefore greater influence diffusion.

The results in Fig. 3 show that empirical influence models not only spread substantially more influence than the baseline models, but also spread more influence than other empirical influence models. This highlights the importance of specifying the correct influence model and thus the correlational structure of the edge propagation thresholds. In other words, misspecification of the correlational structure of the edge propagation thresholds also leads to suboptimal influence spread. It is not surprising that the correct model outperforms the others. What is surprising, however, is the magnitude and economic significance of the misspecification error. The results show that incorrect specification of the correlation structure of edge propagation thresholds can on average lead to 21.7% (95% confidence interval = 19.2–24.2) lower influence for both the IC and LT models compared to the random and the heuristic baselines. The random baseline underestimates influence by 34.0% (95% confidence interval = 29.3–38.7) on average for the LT model and by 17.2% (95% confidence interval = 11.5–22.9) for the IC model. The corresponding numbers for the influence underestimation by the heuristic baseline for the LT model are 23.1% (95% confidence interval = 19.0–27.3) and 12.5% (95% confidence interval = 8.2–16.8) on average for the IC model.

Next, we compare the structural properties of the seeds chosen by the baseline models and the empirical influence models. Figure 4a shows the mean fractional overlap in the seed sets selected by the optimization under different influence models. As can be seen, the overlap in the seed sets is usually quite low (<25%), demonstrating that misspecification errors lead to suboptimal seed selection. The mean overlap between the seeds chosen by empirical influence models compared to the random and heuristic baselines for both the IC and LT models is 19.8% (95% confidence interval = 17.5–22.1). The seed overlap with the random baseline was 11% (95% confidence interval = 8.7–12.8) for the IC model and 21% (95% confidence interval = 16.6–25.5) for the LT model. For the IC model, 89% of the comparisons have an overlap of 25% or less and for the LT model there is an overlap of 30% or less in 61% of the comparisons. Similarly, when compared to the heuristic baseline method (constant for IC and inverse degree for LT), the mean overlap is 15.4% (95% confidence interval = 12.6–18.2) for the IC model and 32% (95% confidence interval = 27.1–36.9) for the LT model. In comparison to heuristic methods, 81% of the comparisons had an overlap of 27% or less and 59% of the comparisons had an overlap of 40% or less.

In some cases, for the LT model, the overlap is higher than average. The higher overlap for the LT model is explained by the fact that it has two parameters, edge propagation probabilities and node-specific thresholds. Our empirical influence models change the edge propagation probabilities based on the eight variants that we describe, but do not interfere with the node-specific thresholds (specified $U(0,1)$), in order to preserve sub-modularity. As influence transmission is determined by both parameters, only one of which is changing, we observe less difference between these models and the baseline results. However, even in cases with higher seed set overlaps, the difference in seed sets is substantial enough to create an economically important difference in the influence spread achieved by the seed nodes chosen by the correct empirical influence model compared to the LT model (as can be seen in Fig. 3).

Although the baseline methods do not choose the same seeds as the empirical influence models, they may be choosing structurally equivalent seeds, that is seeds that display similar structural network characteristics and influence and susceptibility parameters. We therefore compare multiple structural properties of seeds nodes chosen under different influence models, including their degree, Burt's constraint[36] and the Gini coefficient of their influence and susceptibility parameters (Fig. 4b–e).

The seed sets chosen by empirical influence models have lower degrees and higher Burt's constraints on average, compared to the heuristic baselines. This indicates that the structural characteristics

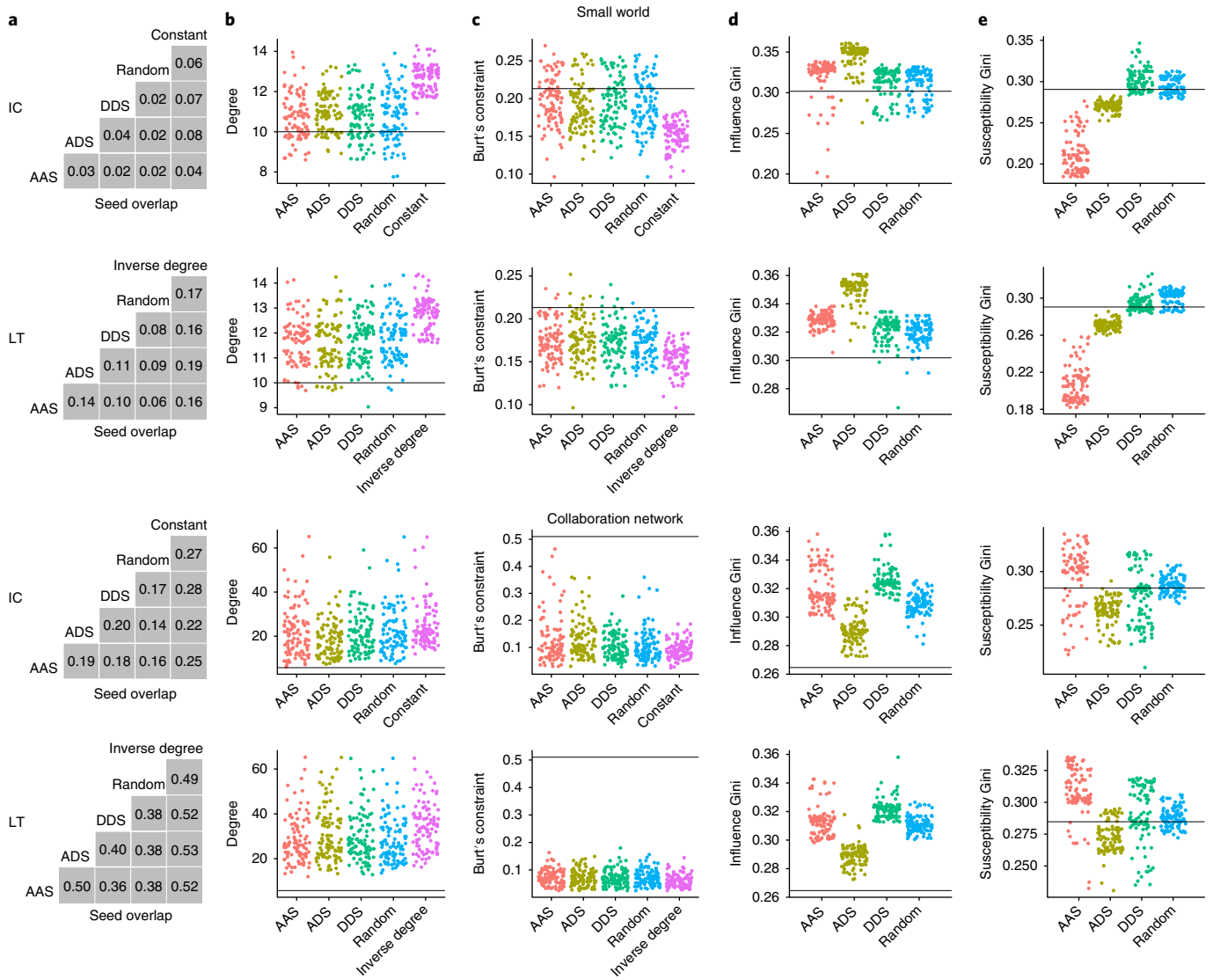

Fig. 4 | Overlap among and structural differences between seed sets under different influence models. Overlap among and structural differences between seed sets chosen by influence maximization under different influence models conditional on the seed set size (the number of initial seed nodes specified by influence maximization) under different influence models (AAS, ADS and DDS) for the IC and LT models in a synthetic small world network (top) and the arXiv high-energy physics collaboration network (bottom). **a**, Mean fractional overlap (averaged over 10 random draws) between the seed sets chosen by the various empirical influence models (AAS, ADS or DDS) and the baseline models (inverse degree, random or constant). For example, the number 0.06 in the top row indicates that only 6% of the seeds were shown to be in common between the constant and the random baseline models for the IC model for the synthetic small world dataset. **b**,**c**, The various structural properties of the chosen seed sets (degree (**b**) and Burt's constraint (**c**)). The horizontal black line shows the mean degree (**d**) and Burt's constraint (**c**) of the entire network. **d**,**e**, The Gini coefficient of the influence (**d**) and susceptibility (**e**) parameters of a seed node and the sub-network induced by its friends-of-friends. The horizontal black line shows the mean influence (**d**) and susceptibility (**e**) Gini for the entire network for the random baseline model.

that define optimal seeds under different influence models are very different. Not only are the seeds chosen by empirical influence models less well-connected than those chosen by the heuristic methods, but they also have more cohesive networks or a greater density of connections among their contacts. Assuming that the cost of convincing a node to broadcast the advertiser's message is proportional to its degree, this finding suggests that seeding nodes that are less well-connected, but that have cohesive, embedded ties with their contacts are more likely to maximize influence diffusion, supporting similar findings that have been described previously[35]. In small world networks, we also see lower Gini coefficients of susceptibility parameters in the AAS models, because these networks have distinct clusters with dense connections within clusters and few connections across clusters. The assortative distribution of susceptibility in these networks creates greater similarity within clusters and thus lower variability in susceptibility across nodes in a given neighbourhood. The degree and Burt's constraint distributions of optimal seeds under empirical influence models are similar to the random baseline model, but actual seeds that are chosen and the implied influence diffusion are very different. This implies that the results are not only driven by the introduction of heterogeneity in propagation thresholds, but also by the specification of the correct correlation structure of those propagation thresholds. It is natural to ask how the choice of opinion dynamics models impacts the seed sets and influence spread. Here, we used the IC and LT models to model opinion dynamics, because they are the most widely studied

models in the broad class of cascade- and threshold-based models. Our findings on the relative performance of the various approaches are consistent across both the IC and LT models, although the absolute performance differs across the IC and LT models as well as across the different datasets, as expected. We hypothesize that our results will hold for other models in the cascade- and threshold-based class of models, because, at a high level, their mechanics are similar.

Recent debates about the usefulness of individual-level psychological and behavioural data (such as the introversion or extraversion of individuals) in micro-targeting campaigns have focused almost exclusively on the effectiveness of such campaigns in changing the opinions or behaviours of the targeted individuals. Our work, however, implies that the use of such empirical data (which may be correlated with individuals' influence and susceptibility) in network seeding could also impact the spread of such behaviours from the targeted individual to their friends, thereby affecting the overall spread of the behaviours and opinions in society.

The influence models that are currently used for influence maximization do not reflect the most recent empirical evidence on how influence diffuses in human social networks. We therefore specified more realistic empirical influence models across twelve commonly used networks in the literature to study how influence model misspecification affects influence diffusion and the optimal seed nodes chosen by influence maximization. The results of our analysis show that ignoring assortativity and the joint distribution of influence and susceptibility leads traditional models to underestimate influence propagation by 21.7% on average for a fixed seed set size. The superior performance of empirical influence models cannot be explained solely by either the incorporation of heterogeneity into the influence distribution in the network or the specification of assortativity in propagation thresholds. Specifying the correct functional forms of heterogeneity and assortativity—whether, for example, influence and susceptibility are assortative or disassortative—is essential to achieving optimal seed selection.

Empirical influence models select optimal nodes that have substantially lower degrees and higher Burt's constraints (or ego network density) compared to heuristic baseline models. However, they have similar degree, centrality and Burt's constraint distributions compared to the random baseline model, suggesting that structural properties alone do not characterize the differences between the chosen seed sets. This highlights the importance of empirically estimating the correct latent influence and susceptibility parameters of nodes in a given network in order to choose the optimal seed sets and indicates that access to behavioural and psychological data is likely to improve influence maximization beyond models that only consider network structure.

Recent empirical advances in using new observational techniques[33] or randomized experiments[37–41] to identify influence and susceptibility in networks provide new opportunities for specifying more accurate, contextual influence models when using influence maximization to identify optimal targets of public policy interventions or business advertising. Our results suggest that the growing body of research on influence maximization needs to incorporate results and insight from the empirical literature on influence identification in order to become more realistic and practically applicable. Furthermore, our empirical models of the non-uniform joint distribution of influence and susceptibility, suggest that social influence is an edge property rather than a node (or individual-specific) property. Individuals experience heterogeneity in their ability to persuade their friends or neighbours. An interesting direction for future research, therefore, is to investigate more realistic influence model specifications, which incorporate context-specific empirical evidence on assortativity and the joint distribution of influence and susceptibility in the specific networks for which influence is being maximized.

Methods

Model specification. Current state-of-the-art influence maximization approaches use simple models of edge propagation probabilities specified as constant, random or inversely proportional to a node's in-degree. Here we assume that an individual's influence and susceptibility are distinct and individually specified. Given a graph $G(V,E)$ connecting a set of V nodes and E edges, such that $|V|=n$, in a binary adjacency matrix which indicates the presence or absence of edges in the graph, we associate two p-dimensional parameters, representing influence $\Lambda_i=\{\lambda_{i_1}, \lambda_{i_2}, \ldots, \lambda_{i_p}\}$ and susceptibility $\Theta_i=\{\theta_{i_1}, \theta_{i_2}, \ldots, \theta_{i_p}\}$, with each node and construct the edge propagation probability for an edge e_{ij} pointing from node i to node j as the normalized inner-product of influence and susceptibility $\frac{\Lambda_i^{\top}\Theta_j}{\|\Lambda_i\|\|\Theta_j\|}$. Each dimension of influence and susceptibility lies between 0 and 1 and could, for instance, represent the influence and susceptibility of that individual for a specific behaviour. Note that in theory, the influence and susceptibility parameters are generally defined as p-dimensional vectors, whereas we assume for our analysis that they are scalars. While we hope that future studies will embrace the dimensionality of influence and susceptibility and explore how variation in influence and susceptibility across behaviours affects influence maximization, such analysis is beyond the scope of the current work. Assuming nodes i and j are connected and that node i has already been activated at time $t-1$, then propagation at time t occurs according to a simple rule: flip a coin with a probability that is equal to the normalized inner-product of the influence of i and the susceptibility to influence of j (that is $\frac{\Lambda_i^{\top}\Theta_j}{\|\Lambda_i\|\|\Theta_j\|}$) such that each activated node gets only one chance to activate each of their non-activated neighbours. This gives rise to non-uniform edge propagation probabilities that are a function of the correlation and assortativity patterns between an individual's own influence and susceptibility and those of their neighbours.

We consider eight specifications of the empirical influence model that vary in (1) the extent to which influence is assortative or disassortative (governing the degree to which influential people associate or dissociate with each other); (2) the extent to which susceptibility is assortative or disassortative (governing the degree to which susceptible people associate or dissociate with each other); and (3) the correlation between individuals' influence and their own susceptibility (governing the degree to which influential people tend to be susceptible to influence). The eight variants of the empirical influence models that we specify are as follows.

(1) Assortative influence, assortative susceptibility, complement influence–susceptibility (AAC): $\rho_{\Lambda_i\Lambda_{j\in N(i)}}>0,\ \rho_{\Theta_i\Theta_{j\in N(i)}}>0,\ \rho_{\Lambda_i\Theta_i}>0\,\forall\, i\in V$

(2) Assortative influence, assortative susceptibility, substitute influence–susceptibility (AAS): $\rho_{\Lambda_i\Lambda_{j\in N(i)}}>0,\ \rho_{\Theta_i\Theta_{j\in N(i)}}>0,\ \rho_{\Lambda_i\Theta_i}\leq 0\,\forall\, i\in V$

(3) Assortative influence, disassortative susceptibility, complement influence–susceptibility (ADC): $\rho_{\Lambda_i\Lambda_{j\in N(i)}}>0,\ \rho_{\Theta_i\Theta_{j\in N(i)}}\leq 0,\ \rho_{\Lambda_i\Theta_i}>0\,\forall\, i\in V$

(4) Assortative influence, disassortative susceptibility, substitute influence–susceptibility (ADS): $\rho_{\Lambda_i\Lambda_{j\in N(i)}}>0,\ \rho_{\Theta_i\Theta_{j\in N(i)}}\leq 0,\ \rho_{\Lambda_i\Theta_i}\leq 0\,\forall\, i\in V$

(5) Disassortative influence, assortative susceptibility, complement influence–susceptibility (DAC): $\rho_{\Lambda_i\Lambda_{j\in N(i)}}\leq 0,\ \rho_{\Theta_i\Theta_{j\in N(i)}}>0,\ \rho_{\Lambda_i\Theta_i}>0\,\forall\, i\in V$

(6) Disassortative influence, assortative susceptibility, substitute influence–susceptibility (DAS): $\rho_{\Lambda_i\Lambda_{j\in N(i)}}\leq 0,\ \rho_{\Theta_i\Theta_{j\in N(i)}}>0,\ \rho_{\Lambda_i\Theta_i}\leq 0\,\forall\, i\in V$

(7) Disassortative influence, disassortative susceptibility, complement influence–susceptibility (DDC): $\rho_{\Lambda_i\Lambda_{j\in N(i)}}\leq 0,\ \rho_{\Theta_i\Theta_{j\in N(i)}}\leq 0,\ \rho_{\Lambda_i\Theta_i}>0\,\forall\, i\in V$

(8) Disassortative influence, disassortative susceptibility, substitute influence–susceptibility (DDS): $\rho_{\Lambda_i\Lambda_{j\in N(i)}}\leq 0,\ \rho_{\Theta_i\Theta_{j\in N(i)}}\leq 0,\ \rho_{\Lambda_i\Theta_i}\leq 0\,\forall\, i\in V$

where ρ_{xy} denotes the Pearson's correlation between x and y and $N(i)$ denotes the set of neighbours of the node i. For example, consider the empirical influence model specification ADC, which entails a positive correlation between a node's influence parameter and their neighbours' influence parameters (assortative); a negative correlation between a node's susceptibility parameter and their neighbours' susceptibility parameters (disassortative); and a positive correlation between their own influence and susceptibility parameters (complementarity). Note that by assortative/disassortative we mean influence or susceptibility are assortative or disassortative, but our framework could be used in future work to denote assortativity or disassortativity in behaviours or 'traits' more generally.

We then quantify the impact of the different correlation and assortativity assumptions in these eight specifications on the outcomes of influence maximization, including the final extent of influence diffusion under empirical influence models compared to current baseline models and the optimal seed sets chosen to maximize influence diffusion under empirical influence models compared to current baseline models (including the network structural differences between the seed sets). We begin with basic IC and LT models into which we incorporate empirically verified influence propagation parameters. We maintain all of the standard assumptions of influence maximization, including greedy optimization, the size of the seed-set k and discrete time dynamics. We do this to

distinguish our contribution from prior work and to ensure that no confounding or co-varying factors can explain our results.

Graph generation and parameterization. Influence models run on parameterized networks with known structure and distributions of influence and susceptibility over nodes. So, we generated synthetic graphs (small world and preferential attachment), collected data on commonly used empirical graphs from the influence maximization literature (for example, collaboration or citation graphs) and performed correlated label propagation on these real and synthetic graphs to generate the influence and susceptibility parameters satisfying the eight empirical influence models described above (see Supplementary Information for details about the algorithm).

The iterative graph labelling procedure extends previously published work[42] and performs an initial binary labelling of the graph corresponding to 'high' and 'low' types for both influence and susceptibility. We generate real values for influence and susceptibility by conditioning on node type (that is, high or low) and then drawing samples from two well-separated Beta distributions, for the influence and susceptibility parameters (see Supplementary Information for details). Attributed graph models[43] are another way of generating graphs with correlated attributes (influence and susceptibility). However, unlike our setting, they generate both graphs and attributes. As a robustness check, the Supplementary Information contains influence maximization results for a setting in which both the graph and the attributes are generated using attributed graph models. The conclusions of the work do not change under this parameterization. Once we established the influence and susceptibility labels, we defined edge propagation probabilities as $\frac{\Lambda_i^{\mathrm{T}}\Theta_j}{\|\Lambda_i\|\|\Theta_j\|}$. The node threshold parameter in the LT model is assumed to be distributed $U(0,1)$ to preserve the sub-modularity of the influence maximization procedure. The proof of the sub-modularity of our empirical influence propagation specifications, which incorporate empirical evidence into IC and LT models, follows previously published studies[3,4] and is presented in the Supplementary Information. Our labelling does not alter the structure of the graphs in any way. It simply labels nodes with influence and susceptibility parameters. In the case of the IC model, the constant baseline model might have more (or less) influence spread than our eight models of empirical influence maximization just by virtue of having more (or less) probability mass on its edges, so we ensure, via normalization, that the total sum of edge weights is the same across all the graphs and that the difference in influence diffusion across models emanates only from the way in which that fixed probability mass is distributed across the network. We do not perform this normalization for the LT model, because it ensures, by design, that each node's incoming weights sum to 1.

Influence maximization and model comparison. Once we generated the graphs and the influence and susceptibility parameters, the influence maximization procedure is straightforward. We use the recently proposed two-phase influence maximization algorithm[21] for influence maximization under the adapted IC and LT models. The number of seeds is set to 100 and the epsilon parameter to 0.1 as has previously been suggested[21]. We compare our empirical influence maximization models with three sets of baseline models that are commonly used in the influence maximization literature[3,16,17]: (1) models that assume randomly distributed influence and susceptibility parameters; (2) models that assume a constant edge propagation probability of 0.1 (called constant) for the IC model; and (3) models that assume an edge propagation probability inversely proportional to the in-degree of the node (called inverse degree) for the LT model.

Code availability. Code for all the models and analyses is available at https://www.dropbox.com/s/iimtqswiesl4skd/inf-max-data-code-release.zip?dl=0.

Data availability. The data that support the findings of this study are available at https://www.dropbox.com/s/iimtqswiesl4skd/inf-max-data-code-release.zip?dl=0.

Received: 1 September 2017; Accepted: 9 April 2018;
Published online: 21 May 2018

References

1. Kitsak, M. et al. Identification of influential spreaders in complex networks. *Nat. Phys.* **6**, 888–893 (2010).
2. Banerjee, A., Chandrasekhar, A., Duflo, E. & Jackson, M. The diffusion of microfinance. *Science* **341**, 1236498 (2013).
3. Kempe, D., Kleinberg, J. & Tardos, É. Maximizing the spread of influence through a social network. In *Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 137–146 (2003).
4. Kempe, D., Kleinberg, J. & Tardos, É. Maximizing the spread of influence through a social network. *Theory Comput.* **11**, 105–147 (2015).
5. Centola, D. & Macy, M. Complex contagions and the weakness of long ties. *Am. J. Sociol.* **113**, 702–734 (2007).
6. Bakshy, E., Rosenn, I., Marlow, C. & Adamic, L. The role of social networks in information diffusion. In *Proc. 21st International Conference on World Wide Web* 519–528 (2012).
7. Centola, D., Eguiluz, V. & Macy, M. Cascade dynamics of complex propagation. *Physica A* **374**, 449–456 (2007).
8. Christakis, N. & Fowler, J. The spread of obesity in a large social network over 32 years. *New Engl. J. Med.* **357**, 370–379 (2007).
9. Domingos, P. & Richardson, M. Mining the network value of customers. In *Proc. 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 57–66 (2001).
10. van den Bulte, C. & Joshi, Y. New product diffusion with influentials and imitators. *Mark. Sci.* **26**, 400–421 (2007).
11. Altarelli, F. et al. Containing epidemic outbreaks by message-passing techniques. *Phys. Rev. X* **4**, 021024 (2014).
12. Newman, M. Spread of epidemic disease on networks. *Phys. Rev. E* **66**, 016128 (2002).
13. Chen, Y. et al. Finding a better immunization strategy. *Phys. Rev. Lett.* **101**, 058701 (2008).
14. Kawachi, I. & Berkman, L. Social ties and mental health. *J. Urban Health* **78**, 458–467 (2001).
15. van Leeuwen, J. *Handbook of Theoretical Computer Science, Vol. A: Algorithms and Complexity* (MIT Press, Cambridge, MA, 1991).
16. Chen, W., Wang, Y. & Yang, S. Efficient influence maximization in social networks. In *Proc. 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 199–208 (2009).
17. Wang, C., Chen, W. & Wang, Y. Scalable influence maximization for independent cascade model in large-scale social networks. *Data Min. Knowl. Discov.* **25**, 545–576 (2012).
18. Borges, C., Brautbar, M., Chayes, J. & Lucier, B. Maximizing social influence in nearly optimal time. In *Proc. 25th Annual ACM-SIAM Symposium on Discrete Algorithms* 946–957 (2014).
19. Leskovec, J. et al. Cost-effective outbreak detection in networks. In *Proc. 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 420–429 (2007).
20. Goyal, A., Lu, W. & Laksmanan, L. Celf++: optimizing the greedy algorithm for influence maximization in social networks. In *Proc. 20th International Conference Companion on World Wide Web* 47–48 (2011).
21. Tang, Y., Xiaokui, X. & Shi, Y. Influence maximization: near-optimal time complexity meets practical efficiency. In *Proc. 2014 ACM SIGMOD International Conference on Management of Data* 75–86 (2014).
22. He, X. & Kempe, D. Robust influence maximization. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 885–894 (2016).
23. Chen, W., Lin, T., Tan, Z., Zhao, M. & Zhou, X. Robust influence maximization. In *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 795–804 (2016).
24. Granovetter, M. Threshold models of collective behavior. *Am. J. Sociol.* **83**, 1420–1443 (1978).
25. Goldenberg, J., Libai, B. & Muller, E. Talk of the network: a complex systems look at the underlying process of word-of-mouth. *Mark. Lett.* **12**, 211–223 (2001).
26. Goldenberg, J., Libai, B. & Muller, E. Using complex systems analysis to advance marketing theory development: modeling heterogeneity effects on new product growth through stochastic cellular automata. *Acad. Mark. Sci. Rev.* **9**, 1–18 (2001).
27. Gomez-Rodriguez, M. et al. Influence estimation and maximization in continuous-time diffusion networks. *ACM Trans. Inf. Syst.* **34**, 9 (2016).
28. Morone, F. & Makse, H. A. Influence maximization in complex networks through optimal percolation. *Nature* **524**, 65–68 (2015).
29. Newman, M. Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208701 (2002).
30. Aral, S., Muchnik, L. & Sundararajan, A. Engineering social contagions: optimal network seeding in the presence of homophily. *Netw. Sci.* **1**, 125–153 (2013).
31. Bramoullé, Y., Djebbari, H. & Fortin, B. Identification of peer effects through social networks. *J. Econ.* **150**, 41–55 (2009).
32. Golub, B. & Jackson, M. O. How homophily affects the speed of learning and best-response dynamics. *Q. J. Econ.* **127**, 1287–1338 (2012).
33. Aral, S., Muchnik, L. & Sundararajan, A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl Acad. Sci. USA* **106**, 21544–21549 (2009).
34. Aral, S. & Walker, D. Identifying influential and susceptible members of social networks. *Science* **337**, 337–341 (2012).
35. Bakshy, E., Hofman, J., Mason, W. & Watts, D. Everyone's an influencer: quantifying influence on twitter. In *Proc. 4th ACM International Conference on Web Search and Data Mining* 65–74 (2011).
36. Burt, R. Structural holes and good ideas. *Am. J. Sociol.* **110**, 349–399 (2004).

37. Aral, S. Commentary-identifying social influence: a comment on opinion leadership and social contagion in new product diffusion. *Mark. Sci.* **30**, 217–223 (2011).
38. Aral, S. & Walker, D. Creating social contagion through viral product design: a randomized trial of peer influence in networks. *Manage. Sci.* **57**, 1623–1639 (2011).
39. Muchnik, L., Aral, S. & Taylor, S. J. Social influence bias: a randomized experiment. *Science* **341**, 647–651 (2013).
40. Bakshy, E., Eckles, D. & Bernstein, M. Designing and deploying online field experiments. In *Proc. 23rd International Conference on World Wide Web* 283–292 (2014).
41. Aral, S. & Walker, D. Tie strength, embeddedness, and social influence: a large-scale networked experiment. *Manage. Sci.* **60**, 1352–1370 (2014).
42. Ugander, J. & Backstrom, L. Balanced label propagation for partitioning massive graphs. In *Proc. 6th ACM International Conference on Web Search and Data Mining* 507–516 (2013).
43. Pfeiffer, J. III et al. Attributed graph models: modeling network structure with correlated attributes. In *Proc. 23rd International Conference on World Wide Web* 831–842 (2014).

Acknowledgements

We thank to D. Eckles for invaluable discussions. S.A. acknowledges funding and support from the NSF (Career Award 0953832). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

S.A. and P.S.D. contributed equally to all parts of the research and writing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-018-0346-z.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to S.A. or P.S.D.

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