Social Networks 34 (2012) 470–480

Contents lists available at SciVerse ScienceDirect

Social Networks

journal homepage: www.elsevier.com/locate/socnet

Influentials, novelty, and social contagion
The viral power of average friends, close communities, and old news☆

Nicholas Harrigan[a,*], Palakorn Achananuparp[b], Ee-Peng Lim[b]

[a] School of Social Sciences, Singapore Management University, 90 Stamford Road, Level 4, Singapore 178903, Singapore
[b] School of Information Systems, Singapore Management University, 80 Stamford Road, Singapore 178902, Singapore

ARTICLE INFO

Keywords:
Social contagion
Subgraphs
Network motifs
Influentials hypothesis
Community structures
Twitter

ABSTRACT

What is the effect of (1) popular individuals, and (2) community structures on the retransmission of socially contagious behavior? We examine a community of Twitter users over a five month period, operationalizing social contagion as 'retweeting', and social structure as the count of subgraphs (small patterns of ties and nodes) between users in the follower/following network.

We find that popular individuals act as 'inefficient hubs' for social contagion: they have limited attention, are overloaded with inputs, and therefore display limited responsiveness to viral messages. We argue this contradicts the 'law of the few' and 'influentials hypothesis'.

We find that community structures, particularly reciprocal ties and certain triadic structures, substantially increase social contagion. This contradicts the theory that communities display lower internal contagion because of the inherent redundancy and lack of novelty of messages within a community. Instead, we speculate that the reasons community structures show increased social contagion are, first, that members of communities have higher similarity (reflecting shared interests and characteristics, increasing the relevance of messages), and second, that communities amplify the social bonding effect of retransmitted messages.

1. Introduction

We know that many forms of social behavior and ideas can be thought of as contagious: participation in strikes, riots, voting or migration; the spread of innovations, fashions, fads, rumors and job advertisements; the forwarding of emails, blog links, status updates or 'wire' news stories; changes in long term behavior such as education, smoking, diet and crime; the solving of collective action problems, such as climate change; and the failure of infrastructure and social institutions, from power grids to judicial and currency systems.

It is also self-evidently true that the social structure of a community will heavily affect the pattern of the spread of a contagious idea: ideas can only pass between individuals who have a relationship, and so the pattern of these relationships will affect the spread of ideas.

But what is the exact relation between social structures – particularly local social structures – and the successful spread of contagious ideas and behaviors? We examine a community of Twitter users over a five month period, studying 'retweeting' behavior on the follower/following network of users. Twitter messages are modeled in a similar way to a disease on a network, with 'infections' (tweets) being attributes that are passed along the follower/following network, and retweeting another user's message being a sign of infection.

We focus on the effect of two types of social structures: (1) popular individuals (such as users with a large number of followers) and (2) community structures (such as mutual and triadic structures between users). We operationalize the concepts of popular individuals and community structures by counting subgraphs – small patterns of ties and nodes also called graph statistics or network motifs – in the networks of senders and receivers of contagious messages. Subgraphs measuring popular individuals included the number of followers, number following, and number of messages sent (tweets).[1] Subgraphs measuring community structures included the mutual dyad (reciprocity), and a range of triadic structures, including the 3-cycle and the transitive triad.

☆ This research was supported by Singapore Management University Internal Grant C242/MSS9S014, and by a post-doctoral position at Nuffield College, University of Oxford, under the supervision of Peter Hedström. This research was supported by the Singapore National Research Foundation under its International Research Centre@ Singapore Funding Initiative and administered by the IDM Programme Office.

* Corresponding author. Tel.: +65 6828 0842; fax: +65 6828 0423.
E-mail addresses: nharrigan@smu.edu.sg, nickharrigan@gmail.com (N. Harrigan), palakorna@smu.edu.sg (P. Achananuparp), eplim@smu.edu.sg (E.-P. Lim).

0378-8733/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.socnet.2012.02.005

[1] Note that the word 'popularity' is used in this paper to refer to both nodes with high in-degree (the traditional meaning) and nodes with high out-degree (traditionally called 'activity').

We test a range of competing theories of the effect of social structure on contagious ideas. For popular individuals, we test the competing theories of efficient hubs (such as the 'influentials hypothesis' and the 'law of the few' Katz and Lazarfeld, 1955; Merton, 1968; Gladwell, 2000; Rosen, 2000; Watts and Dodds, 2007; Watts, 2007) and inefficient hubs (Barabasi, 2002). For community structures, we test two schools of thought: firstly, those that emphasize the negative effect of community structures on contagion, with a particular focus on theories of the negative effect of message redundancy (and lack of novelty) on contagion (Granovetter, 1973, 1983; Burt, 1992, 2005; Cha et al., 2010; Hill et al., 2006; Leskovec et al., 2007; Wu and Huberman, 2007); and secondly, those that emphasize the positive effect of community structures on contagion, with a focus on theories of the increased user similarity and increased social bonding effect within a community.

This paper begins with a literature review and an outline of our hypotheses (Section 2), followed by an overview of our dataset (Section 3) and methods (Section 4), particularly those methods which involve the counting of subgraphs. We then present our results (Section 5) and a discussion (Section 6) of our findings in light of existing research.

2. Literature review and hypotheses

So what do previous studies argue are the major effects of (1) popularity, and (2) community structures, on the spread of contagious ideas and behaviors? We organize this literature review by major competing theories (explanations), grouping previous studies with the theories they support.

2.1. Popularity

2.1.1. Efficient hubs

A large number of authors emphasize the positive 'influential' role played by highly popular individuals – those with a large number of friends or followers, or a high communication volume – in the spread of socially contagious phenomena. These individuals are presented as analogous to communications hubs that become more efficient and influential as they gain more network partners.

This theory is exemplified by the 'law of the few' and the roles of 'connector' and 'maven' in Gladwell's famous book *The Tipping Point* (Gladwell, 2000). This 'influential's hypotheses', as Watts calls it (Watts and Dodds, 2007; Watts, 2007), has a long tradition in sociological and marketing theory: highly connected individuals have been called variously "opinion leaders" (Katz and Lazarfeld, 1955), "influentials" (Merton, 1968), or "hubs" (Rosen, 2000). Certain recent academic studies have provided limited quantitative support for this perspective. Cha et al. (2010) found in a study of Twitter that content aggregation services and news sites (which had large numbers of followers) were the most retweeted users. The same study found that the higher the status of a user, the greater their likelihood of being 'mentioned' (replied to).

At the theoretically level, Watts argues (Watts and Dodds, 2007; Watts, 2007) and we agree, that the exact mechanisms driving the supposed positive relationship between popularity and social contagion are, at best, poorly specified in the above literature. Nonetheless, we believe that a theory can be knitted together that captures the major elements common to these accounts: what unites these theories of 'efficient hubs' is, firstly, a claim that there is a positive feedback cycle that is reinforcing the importance of popular actors. The exact mechanisms driving positive feedback cycles will varies across cases, but an archetypical example might include (a) greater contacts leading to: (b) better information, (c) greater visibility, (d) increased social status, and (e) increased trust. With each of (b) through (e) directly and/or indirectly leading to (a) greater contacts, the positive feedback cycle is completed. This generates a 'rich get richer' (Barabasi and Albert, 1999) scenario where a very small number of actors monopolize a disproportionate amount of the social contacts, information, visibility, status, and trust in a network.

The second claim of theories of efficient hubs is that these highly connected (popular) actors are the central conduits of contagious information spread. Depending on the particular author/study, this increased role in contagion is a result of (1) their increased number of contacts, and/or (2) their disproportionate possession of the means to influence any one individual–they may be a more trusted source, of higher status, have greater persuasive powers. As a result of both disproportionate friendships and means of influence, a message sent by a popular individual will, on average, be passed-on (i.e. spread, retweeted, forwarded, etc.) to a considerably larger audience than any message received by an average user. This is the theory of 'efficient hubs'.

2.1.2. Inefficient hubs

Another, more critical, section of the literature on social contagion emphasizes that highly connected individuals tend to show sub-optimal viral reproduction rates.[2] In a theoretical study, Golub and Jackson (2007) found that efficient diffusion of influence through a network is limited by the presence of highly influential, high degree nodes. Empirical studies of Twitter have similarly found that the most 'influential' actors – measured by their impact on hashtag adoption – were moderately 'sized' (followers, volume of tweets) users (Yang and Leskovec, 2011). An empirical study of an online bookstore found similar trends: high degree nodes have lower influence per recommendation; nodes tend to only be influential over their close friends; and books with higher numbers of recommendations tend to have lower viral reproduction (Leskovec et al., 2007).

While theoretically these empirical accounts differ in their hypothesized mechanisms, there is a common element to most studies: theories of inefficient hubs tend to emphasize the limits on efficient (1) growth and (2) activity which social hubs face. Limits on (1) growth occur because ties themselves are not costless: forming friendships requires time, and time is necessarily limited. Limits on (2) activity occur because communication itself is not costless: maintaining friendships requires time and energy as well. While in some non-human networks (such as hyperlinks on the internet), ties are costless, in the real human social world this is often not the case: limited time and attention place natural limits on the capacity of individual humans to meaningfully expand their social networks and to be the fulcrums of the spread of socially contagions behavior (Barabasi, 2002).

The theory of 'inefficient hubs' says that those individuals who deviate significantly from the average – who (for whatever reason) push the natural boundaries on the size of a human's social networks – will face the problem of overload and inefficiency. Because of this overload and inefficiency, highly popular individuals are expected to have a disproportionately lesser impact on social contagion than their number of social contacts would suggest. Instead, the theory of inefficient hubs predicts that real socially contagious transmission will occur amongst less 'overloaded', more 'average', individuals.

[2] Watts and Dodds (2007) provide another criticism of the influentials hypothesis: they find that they can simply not simulate a world in which influentials are important in starting contagious outbreaks. They vary a large number of parameters, but find very few situations in which influentials are substantially more influential than the average individual.

2.1.3. Operationalizing popularity

To test the relative merits of the theories of efficient and inefficient hubs, we put forward the following hypothesis:

H1. That popular users will retweet more.

We operationalize the concept of 'popular individuals' with three measures: 'following', 'followers', and 'total tweets'.

2.2. Community structure

2.2.1. Redundancy of messages within a community

There is a long tradition of literature on social contagion that emphasizes the negative role played by 'community structure' in the spread of ideas. At the heart of this literature is the argument that messages passed within a community will tend to be redundant, and therefore lack novelty. Such lack of novelty lowers the incentive of senders to spread messages[3], and lowers the interest of recipients in receiving such messages.

It is important to acknowledge that there are a number of related, but distinct mechanisms linking (1) community structure, (2) redundancy of messages, and (3) social contagion. It is beyond the scope of this paper to completely disaggregate all of the potential processes. Instead the process will be reduced to three separate claims: (i) community structure is likely to lead to messages being redundant (i.e. already received); (ii) such redundancy reduces the novelty of messages sent and received via ties that are part of community structures; and (iii) such lack of novelty reduces the likelihood of sending and receiving such messages (contagion).

Claims (i) and (ii) are most famously articulated in the existing literature by the separate work of Granovetter (1973, 1983) and Burt (1992, 2005). Granovetter's 'strength of weak ties' (1973, 1983) emphasizes that strong ties tend to exist in triads, and triads tend to act like echo-chambers, reproducing only redundant messages. While weak ties might bring new information (such as job opportunities), strong ties (and triadic ties generally) do not. Another famous theory within this literature is Burt's theory of structural holes (1992, 2005). According to Burt, actors whose neighbors are themselves not tied to each other tend to accrue novel information and systematic benefits (brokerage). A corollary of this is that actors within triads – community structures – tend not to be the recipients of (new) messages and information.

Claim (iii) – that novelty drives message contagion – has received a great deal of empirical support from a number of recent studies of message diffusion. Cha et al. (2010) found that the most retweeted users were content aggregation services (Mashable, TwitterTips) and news sites (NYT, The Onion), and hypothesized that this was precisely because of the novelty of the content (i.e. lack of redundancy) of their posts. Macskassy and Michelson (2011) found that individuals' probability of retweeting a message on Twitter was lower when either the individual message, or the user who sent the message, was similar to the user (on a multidimensional subject/topic scale). In short they found that users retweeted messages that were dissimilar from their normal postings, and from users who had dissimilar interests to their own. Similarly, Hill et al. (2006) found that buzzy products (such as new mobile phones) generated much greater sales through social contagion than not-so-buzzy products (such as new pricing plans). In addition, Leskovec et al. (2007) found that popular, cheaper books (widely known, and therefore lacking novelty) have lower viral transmission, while expensive books in niche areas (and thus unknown and containing inherent novelty) have much higher viral transmission. Finally, Wu and Huberman (2007) found that content novelty drives 'attention' on Digg.

2.2.2. Increased similarity within a community

We identify two sets of theories that predict that community structures will have a positive effect on likelihood of social contagion: theories of similarity and theories of social bonding.

Theories of increased similarity within a community tend to argue that common friendships within community structures either (1) induce similarity between individuals, or (2) reflect (latent/unmeasured) similarity between individuals. Such similarity, it is argued, increases message retransmission through a variety of potential mechanisms, including increased relevance of messages, and increased attention paid to individuals who are similar to one's self.

A notable example of the role of similarity in message retransmission is Johnson Brown and Reingen (1987). They found that strong ties were more likely to transmit information because they were more likely to be activated and more likely to be perceived as important. Similarly, Leskovec et al. (2007) found that successful cascades of recommendations of books were more likely to occur within tight-knit communities who shared specific interests or common institutions (for example, the readers of technical or religious books).

2.2.3. Social bonding motive within a community

We put forward what we think is a new theory of the relationship between community structure and social contagion. This 'theory of social bonding' argues that communities encourage message retransmission as a form of social bonding.

Traditional theories of social contagion tend to think the purpose of resending messages is to spread information. There is however a strong argument that resending messages is a way of signaling respect, solidarity, common values and shared social identity. We believe the social bonding motive may be of growing importance with the advent of online social media: the highly public nature of online social media means that the resending of messages is much more than a dyadic act.

A classic case of social contagion as social bonding is 'liking' and 'sharing' on Facebook. Within Facebook, there is a hierarchy of positive responses to a wall post: a 'like', a 'comment', and a 'share'. Reposting a friend's post on one's wall (known as 'sharing') is one of the highest complements in Facebook etiquette and therefore a strong form of social bonding. In a similar way, community structures may actually increase message retransmission because of the stronger bonding effect of retransmission within a larger group. We are unaware of previous studies that have attempted to quantify the contribution of the social bonding motive to message contagion.

2.2.4. Operationalizing community structures

To test these competing theories of community structures, we put forward a series of hypotheses. To put forward these hypotheses requires an operationalization of community structure. We provide a full outline of our method of operationalizing community structure later in the paper, in Section 4.1. Before reading further, the reader may wish to familiarize themselves with Section 4.1 and the associated illustration of the subgraphs themselves in Fig. 1.

In our descriptions of the hypotheses we will refer to three actors: the original sender (or sender) named 'I'; the potential retweeter (or retweeter) named 'J'; and the third party named 'K'.

We begin operationalizing community structure at the dyadic level, using mutual (also called reciprocal) ties. If two nodes both follow and are followed by each other then they are said to have

[3] Providing new information may bring status, social approval, and even financial returns. Sharing a message also has a much lower downside risk if you share few contacts with the original sender: the social distance between the origin and final destination provides a guarantee that the message is unlikely to have the appearance of spam or dated material.

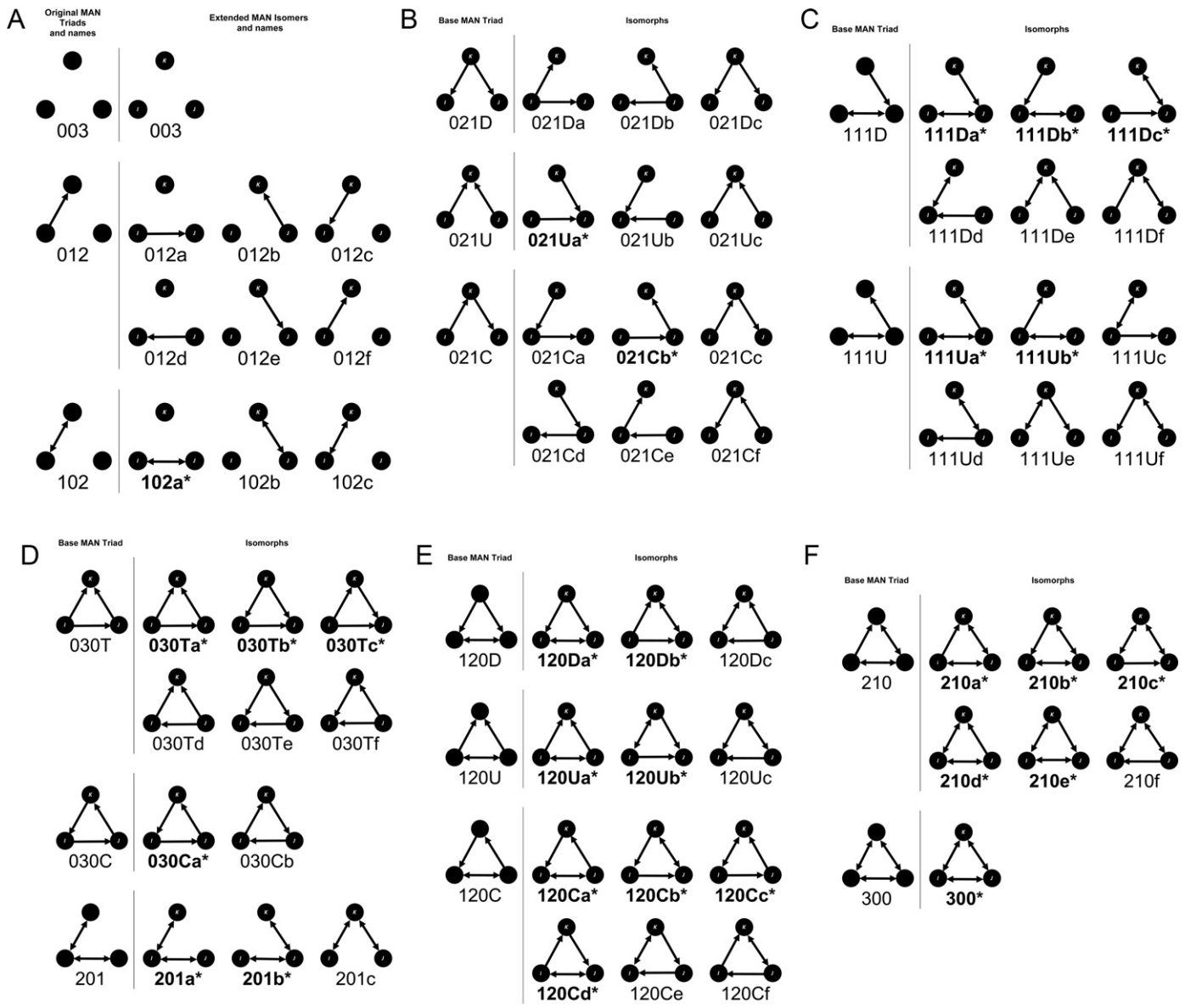

Fig. 1. (Panels A–F) Extended MAN naming system for all isomers on fixed set of nodes. *Subgraphs included in Twitter modeling.

a mutual tie (we can also say that one tie is reciprocated by the other).

H2a. That potential retweeters with reciprocal ties to the original sender will be more likely to retweet.

We operationalize hypothesis 2a by examining the count of mutual ties (102a) between the sender and the potential retweeter. Hypothesis 2a is also partially operationalized in a large number of other subgraphs which incorporate a mutual tie between the sender ('I') and the potential retweeter ('J') such as: 111Da, 120Da, 120Ua, 210a, 300, and many others.

A second dyadic level hypothesis is that potential retweeters (J) with mutual ties to third parties (K) will have a higher likelihood of retweeting.

H2b. That potential retweeters with reciprocal ties to third parties will be more likely to retweet.

Hypothesis 2b is operationalized in 'count of mutual friends' (111Dc). 'Count of mutual friends' (111Dc) reflects the relatively greater embeddedness of the potential retweeter in closer relationships in general. As with hypothesis 2a, hypothesis 2b is also partially operationalized in a large number of other subgraphs which incorporate a mutual tie between the potential retweeter ('J') and their followers ('K'), such as: 201b, 120Db, 120Cb, 210a, 210c, and many others.

The second way we operationalize community structure is by examining the effect of triadic structures: structures where the sender ('I') and potential retweeter ('J') having ties to common third parties ('K'). First we put forward the general hypothesis:

H3. That potential retweeters ('J') and original senders ('I') with ties to common third parties ('K') will be more likely to retweet.

Hypotheses 3 is simply predicts that individuals ('J') embedded in triadic structures ('I' is tied to 'J', 'J' to 'K', and 'K' to 'I') will show higher rates of retweeting.

We disaggregate hypotheses 3 into a series of related claims. In particular, we divided these group structures into a variety of 'types'. We call the first of these types of group structures 'egalitarian groups':

H3a. That potential retweeters ('J') embedded in 'egalitarian groups' will retweet more.

We operationalize the variable 'egalitarian group' with the measure '030Ca'[4] – what is commonly known as the three cycle.

We call the second type a 'group of disciples'. It involves senders and receivers who are both following similar third parties:

H3b. That potential retweeters ('J') embedded in 'groups' where the original sender ('I') and potential retweeter ('J') both follow similar third parties ('K') will retweet more.

We operationalize the variable 'group of disciples' with the measure '030Tb'.

We call the third type a 'group of leaders'. It involves senders and receivers who are both followed by similar third parties:

H3c. That potential retweeters ('J') embedded in 'groups' where the original sender ('I') and potential retweeter ('J') are both followed by similar third parties ('K') will retweet more.

We operationalize the variable 'group of leaders' with the measure '030Ta'.

We call the fourth type of group a 'hierarchical group':

H3d. That potential retweeters ('J') embedded in 'groups' where the potential retweeter ('J') follows third parties ('K') who follow the original sender ('I') will retweet more.

We operationalize the concept of the 'hierarchical group' with the measure '030Tc'.

We call the fifth type of group the 'close community':

H3e. That potential retweeters ('J') embedded in 'groups' where the original sender ('I') and the potential retweeter ('J') both have a mutual ties (reciprocal follower-following) to a common third party ('K') will retweet more.

We operationalize the concept of 'close community', with the measure '210c'.

As with hypotheses 2a and 2b, each of these hypotheses 3a–3e is also partially operationalized in a large number of other subgraphs that incorporate each basic subgraph into a higher order subgraph (i.e. more complex subgraph). For example, 030Ta operationalizes 'group of leaders', but subgraphs 120Db and 120Cc both contain 030Ta within them, and thus 120Db and 120Cc each partially operationalize the concept of 'group of leaders'.

3. Data

We studied a population of 2717 Twitter users. This was the entire population of Twitter users in late November 2009 who listed their location as Singapore, whose profiles were 'public' (i.e. their tweets could be viewed by anyone on the web), and whom met our criteria for being 'real individuals'. Since the aim was to study networks at the individual rather than organizational scale, we took measures to remove from large corporations, news organizations and spammers. To be classified as 'real individuals' (and included in our dataset) users had to: (1) have at least one follower, and (2) not have more than 500 followers.

Our proxy for social contagion was the retweeting of messages. Retweeting is a process similar to forwarding as a user receives a message and then reposts the message to all their contacts (followers). As with standard Twitter messages, retweets go to all followers (unlike email, which sends messages to a select group of contacts).

We define the 'depth' of a retweet as the number of users a message has passed through: an original twitter post is depth = 0; a message retweeted for the first time is depth = 1; if that message is then retweeted by the recipients of the retweeted message, then the message has a depth = 2, etc. Limitations on our data collection meant that we could only be absolutely certain of the chain of retweeting for messages of depth 1. Thus we discarded all messages of depth greater than 1.

Our population retweeted 4440 messages a total of 5090 times (some messages were retweeted by more than one user) in the five months of the study (December 2009–April 2010). Retweeting was a widespread phenomenon: 1736 (63%) of users sent messages that were retweeted by another user, and 1490 (55%) of users retweeted at least one message.

The network on which the contagion took place was the follower-following network of the 2717 Twitter users. Ties were directed, pointing in the direction of information flow: from those followed to those following.

4. Methods

4.1. Subgraph theory and methods

We measure social structure by counting 'subgraphs' in the network surrounding the original senders and potential retweeters. Subgraphs are small configurations (or 'motifs') of ties and nodes. Subgraphs generally contain between 2 and 10 nodes. Examples include the asymmetric dyad (a tie in one direction), the mutual dyad (a reciprocated tie), and a 3-cycle (three nodes, each with one intie and one outtie, pointing in a circle).

Subgraphs provide a number of important advantages over other traditional measures of social network structure, such as centrality (Freeman, 1979) or structurally equivalent blocks (White et al., 1976). First, subgraphs potentially model *local* mechanisms. This makes subgraphs plausible candidates for explanations in conditions where an actor's awareness and rationality are bounded. Second, subgraphs measure the *simplest* forms of network structure. This makes subgraph models parsimonious, and ensures their network models are built on the fewest possible assumptions.

Those wishing to use subgraphs to statistically model social networks face a choice between two major modeling techniques: the subgraph census (formalized in Holland and Leinhardt, 1970) and the logit model (formalized in Holland and Leinhardt, 1981). The subgraph census approach involves the comparison of counts of subgraphs in an observed social network with the distribution of subgraph counts in a set of randomly generated graphs. More advanced conditional random graph distributions control for the outdegrees (U|{Xi+}), the dyad census ((U|MAN) (Wasserman, 1975), and in more recently, all lower order triadic effects (Milo et al., 2002).

The logit model approach to the statistical modeling of subgraphs involves the parameterization of subgraphs as explanatory variables in a logistic regression, where the outcome variable is the formation (or dissolution) of a single tie or the change in a single attribute. Advances in logit subgraph models have involved the expansion of the complexity of subgraphs (Frank and Strauss, 1986; Hunter, 2007; Robins et al., 2001a,b, 2006; Snijders et al., 2006; Wasserman and Pattison, 1996), the improvement of estimation techniques (Snijders, 2002), and the extension to models of social influence (Robins et al., 2001a,b) and co-evolution of networks and attributes (Snijders et al., 2007).

We adopt the logit model approach of Robins et al. (2001a,b)'s: the logit social influence model. The dependent variable is an attribute change: whether the recipient retweets the message (1) or not (0). The independent variables are subgraphs based on a modified version of the MAN naming system for isomorphic triads (Wasserman, 1975: 4–5). The original MAN naming convention, listed in column 1 of Fig. 1, involves assigning three numbers and potentially a trailing letter to each subgraph configuration. The first

[4] For an explanation of the subgraph nomenclature (for example, 030Ca and 030Tb) see Section 4.1 and Fig. 1.

number is the number of mutual ties (M). The second is the number of asymmetric ties (A). The third is the number of null ties (N). The trailing letters further distinguish between configurations: up (U), down (D), cyclic (C) and transitive (T).

The original MAN naming convention is insufficient for our study. The problem is that it does not differentiate between all isomers (64 in total) that occur when the identities of the three nodes in a triad are distinguished (for example by naming them 'I', 'J', 'K' or by numbering them 1, 2, 3). As mentioned earlier, our study distinguishes between the identities of the nodes because one node is the 'original sender' ('I'), another node the 'potential retweeter' ('J'), and a third node is the 'third party' ('K'). These identities impose distinctions between triad isomers that the original MAN conventions aggregate. For example, in Fig. 1 (Panel C) 111Da and 111Db have very different meanings in our Twitter modeling: 111Da is the count of the followed of *potential retweeters* who have a mutual tie with the sender, while 111Db is the count of the followed of *senders* who have a mutual tie with the potential retweeter. Aggregation of such isomers into one count (such as aggregating 111Da, 111Db and 111Dc into one measure 111D) would generate variables with unclear meaning and make interpretation of subsequent results difficult, if not impossible.

The modified MAN naming system is outlined in full in columns 2, 3, and 4 in Fig. 1 (Panels A–F). The first three numbers and any trailing upper-case letter identify the base MAN triad from which the isomer originates. If there is more than one isomer of the MAN triad, then these are differentiated by labeling with a final lower-case letter. Labeling of isomers starts at 'a' and goes as high as 'f' in cases where there are six isomers of the MAN triad.

Amongst the isomers in the modified MAN naming system (columns 2–4 in Fig. 1) a distinction is drawn between those isomers included in the Twitter models (bolded with an asterisk (*)), and those that are excluded. There are two reasons why isomers are excluded from our Twitter modeling. First, if an isomer lacks an x_{ij} tie (sender-potential retweeter tie) then the count for these subgraphs will automatically be zero. This is because it is impossible for actor I to send the original tweet to actor J if actor J is not a follower of actor I. Subgraphs excluded for this reason include 003, 012b, and 021Ub. The second reason why some isomers are excluded from our Twitter modeling is because some isomers always takes the same value across all potential retweeters (actor J) of the same message (from actor I). If the count of a subgraph (isomer) does not vary across receivers, then it cannot explain relative differences in receivers' propensity to retweet. Three isomers are excluded because of this characteristic: 012a, 021Da, and 021Ca.

4.2. Regression model

We use a conditional logistic regression model (McFadden, 1973). The dependent variable is a binary attribute variable: did the receiver of the message retweet or not? Observations (cases) are all ***potential*** retweeters of messages: all recipients of a message that was eventually retweeted by at least one user. Observations are grouped by message and thus the estimates capture within-group differences.

The primary strength of the conditional logit model (also called a 'fixed-effect model') is that it removes biases created by unobserved heterogeneity. In the case of Twitter, the unobserved heterogeneity that is removed includes: the message content (since we are only comparing between recipients of the same message), the senders characteristics (since we are only comparing between recipients who are following the same sender), and the senders' friends and network characteristics (since we are comparing between people who are following the same sender). Thus, when interpreting the results of a conditional logit model we do not need to take account of possible biases generated by the message content or sender characteristics of messages that are retweeted. This advantage of the conditional logit model is also demonstrated by its widespread use across social network studies (for example, Snijders et al., 2010; Helleringer and Kohler, 2005).

The model is given algebraic form in Eq. (1). $\boldsymbol{j_i}$ is a vector of the binary attribute variable 'retweet' of all recipients j (potential retweeters) of all messages i. $\boldsymbol{j_i}$ is equal to 1 if actor j retweeted message i, and zero otherwise. $\boldsymbol{X}$ is the follower-following network of all twitter users in our dataset. $\boldsymbol{Z}$ is the actor attributes (in this case there are only two actor attributes: 'retweet' ($\boldsymbol{j_i}$, the dependent variable) and 'number of tweets' (a covariate)). A_i is the set of all recipients of the particular message i. b_i is an actor from the set A_i. β_k is a vector of parameters to be estimated across all cases. $s_{j_i k}$ is the vector of subgraph counts for actor $\boldsymbol{j_i}$. The subscript k labels each distinct parameter/subgraph in the model (for example, mutual dyads may be $k = 1$; 3-cycles may be $k = 2$; etc.).

Our model aims to show that *actors who do retweet messages* (i.e. where $\boldsymbol{j_i} = 1$) have systematically different subgraphs (s_j) from the subgraphs (s_b) surrounding actors as a whole (A_i).

A conditional logit model estimates the effects of covariates (in this case subgraphs $(s_1, \ldots, s_k)$) on an outcome (in this case retweeting ($\boldsymbol{j_i}$)) by estimating a set of parameters $(\beta_1, \ldots, \beta_k)$ that maximizes the log of Eq. (1). The numerator of Eq. (1) is the odds of j retweeting, while the denominator is the sum of the odds of all actors (A_i) retweeting.

$$Pr(\boldsymbol{j_i} = 1 | \boldsymbol{X}, \boldsymbol{Z}) = \frac{\exp\left(\sum_k \beta_k s_{\boldsymbol{j_i} k}\right)}{\sum_{b_i \in A_i} \exp\left(\sum_k \beta_k s_{b_i k}\right)} \quad (1)$$

The parameters $(\beta_1, \ldots, \beta_k)$ are straight forward to interpret. They represent the effect of subgraphs $(s_1, \ldots, s_k)$ on the likelihood of retweeting, over and above any general increase in the likelihood of retweeting caused by characteristics of either the sender or the message (i) itself.

5. Results

Table 1 provides a summary of the univariate statistics of the dataset. 2.7% of recipients in the dataset retweeted the message they received.[5] 67% of recipients were 'followed' by the original sender (a mutual tie). During the five month period, the average number of total messages (including all tweets and retweets) sent by each recipient was 164. Total messages showed considerable variation across users (standard deviation 147, maximum 1644). Users were, on average, followed by 61 people, and following 50 people. Again this showed considerable variation across users (standard deviation ~70, maximum ~400–500).

To allow easy comparison of the major classes of our MAN triad isomers, we divide the graph statistics in Table 1 into four rough categories: basic stars, complex stars, basic triads and complex triads.

The counts for these graph statistics vary considerably, but do follow a pattern: simpler graph statistics are more common, more complex graph statistics are rarer. On average, users are embedded in 35–70 of each of the 'basic star' (UMN numbers 111) configurations. Again, there is considerable variation across users (standard deviation ~50–100, maximum ~350–500). On average, users are embedded in approximately 30–50 'complex star' (UMN numbers 201) configurations, again with considerable variation across users.

[5] Remember that the observations (cases) were all ***potential*** retweeters of messages: all recipients of a message that was eventually retweeted by at least one user. Those users who received a tweet which was not retweeted by any follower were excluded from the analysis.

Table 1
Univariate statistics for variables (observations are those used in model (N = 115,986)).

Variable	Mean	SD	Min	Max
Retweet	0.027	0.16	0	1
Mutual (102a)	0.67	0.47	0	1
Total tweets	164	147.1	0	1644
Followers (021Cb)	60.61	71.45	1	500
Following (021Ua)	50.14	66.42	1	414
Basic stars				
111 Da	41.27	66.34	0	413
111Db	73.43	102.23	0	399
111Dc	37.47	51.28	0	398
111Ua	38.19	65.33	0	529
111Ub	68.38	86.24	0	350
Complex stars				
201a	49.54	64.12	0	236
201b	28.44	50.02	0	397
Basic triads				
030Ca	13.62	15.84	0	189
030Ta	16.29	17.74	0	223
030Tb	16.51	17.88	0	183
030Tc	16.29	17.06	0	196
Complex triads				
120Ca	11.43	16.27	0	189
120Cb	11.47	13.80	0	176
120Cc	12.00	13.71	0	172
120Cd	12.00	16.86	0	196
120Da	12.76	18.14	0	183
120Db	12.74	14.50	0	193
120Ua	13.04	18.07	0	223
120Ub	13.04	14.00	0	159
210a	10.05	14.55	0	193
210b	9.53	14.08	0	176
210c	10.05	11.90	0	157
210d	10.02	14.08	0	172
210e	10.02	14.15	0	159
300	8.29	12.12	0	157

Table 2
Bivariate statistics as Beta coefficients in a conditional logit model.

	Beta coefficient	z-Statistic
Basic stars		
111Da	0.0008	0.58
111Db	−0.003*	−2.52
111Dc	0.02***	5.89
111Ua	−0.004*	−2.29
111Ub	−0.001	−1.07
Complex stars		
201a	−0.003	−1.82
201b	0.004*	2.03
Basic triads		
030Ca	0.04***	13.52
030Ta	0.04***	14.63
030Tb	0.04***	13.59
030Tc	0.04***	14.68
Complex triads		
120Ca	0.04***	11.63
120Cb	0.05***	14.40
120Cc	0.05***	14.13
120Cd	0.04***	12.29
120Da	0.04***	11.69
120Db	0.05***	16.32
120Ua	0.03***	12.27
120Ub	0.05***	13.91
210a	0.04***	13.71
210b	0.04***	12.44
210c	0.06***	15.31
210d	0.04***	12.20
210e	0.04***	11.76
300	0.05***	13.26

Each model also includes (not reported here) the four simplest graph statistics as controls: mutual (102a), total messages, followers (021Cb) and following (021Ua).
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

'Basic triads' (UMN numbers 030) are considerably less prevalent. On average, each user is embedded in approximately 13–16 of each of these configurations, but, as with all other statistics, there is considerable variation (standard deviation ∼17, maximum ∼200). 'Complex triads' (UMN numbers 120, 210 and 300) are rarer still. On average, each user is embedded in approximately 8–13 complex triads, with substantial variation across users (standard deviation ∼14, maximum∼150–200).

Table 2 presents bivariate statistics as coefficients in a conditional logit model. This is presented as a partial substitute for a correlation matrix because in a simple correlation matrix the (largely positive) effects of higher order graph statistics are swamped by the (largely negative) effects of the simple graph statistics (mutual, total tweets, followers, following).

Estimated coefficients are reported as odds ratios for each of the individual configurations. Each model also includes (not reported here) the four simplest graph statistics as controls: mutual, total messages, followers and following.

There are three major results worthy of note in Table 2. First, all but one of the star statistics – both complex and simple – have little (*$p < 0.05$) or no significant ($p > 0.05$) association with retweeting. Second, the exception to this result is configuration 111Dc, which has a positive effect on probability of retweeting. In the presence of controls, 111Dc is a measure of the effect of *the potential retweeter's mutual ties to third parties*. In other words, individuals with many third party 'friends' who they have mutual ties to are more likely to retweet.

Third, all of the triadic configurations have a positive significant association with retweeting (***$p < 0.001$, z-statistics > 10). The magnitude of this is significant and substantial: one higher order triadic configuration difference in two individuals networks results in a difference in odds of retweeting of approximately 100:103, i.e. for every hundred times an individual without a higher order triadic configuration retweets, a person with one of these configurations will retweet 103 times. While this may seem small, the relatively high mean values for these configurations (∼8–16), and their substantial variation across individuals (standard deviation ∼13–17) leaves considerable scope for these variables to have a substantial effect on retweeting behavior.

Table 3 reports the results of six models. The first five models (Models 1–5) are conditional logistic regressions with no forwards or backwards selection: the variables included in each model are those with parameter estimates listed in Table 3. The sixth model included all modified MAN variables, plus the four basic variables (total tweets, mutual, followers, following). The sixth model was simplified using forwards selection of variables, with a 0.05 significance level for addition to the model. The four simple variables (mutual, total tweets, followers, following) were locked into Model 6.

Rather than report on the results of each model separately, we report results by configuration. Total tweets had a highly significant and negative association with propensity to retweet in all six models ($p < 0.001$, beta = −0.0005). A mutual tie (102a) between the sender and potential retweeter had a highly significant and positive association with retweeting in all models ($p < 0.001$, beta = ∼1). Followers (021Cb) had a strong negative association with retweeting in Model 2 ($z = 17$), but this effect largely disappeared ($z = -0.89$) once the number following (021Ua) was included in Model 3. Following (021Ua) was the single most powerful explanatory variable in the models. The more users a particular actor follows, the lower the propensity to retweet ($p < 0.001$, beta = ∼−0.01).

The triadic configurations 030Ta and 120Db had a strong positive and significant relationship with the propensity to

Table 3
Conditional logit models of propensity to retweet.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Total tweets	−0.0005*** (−3.91)	−0.0005*** (−3.73)	−0.0005*** (−3.61)	−0.0005*** (−3.84)	−0.0005*** (−3.79)	−0.0005*** (−3.78)
Mutual Tie (102a)	1.28*** (21.65)	1.33*** (22.18)	1.17*** (18.88)	0.94*** (14.91)	1.18*** (13.49)	1.10*** (12.23)
Followers (021Cb)		−0.01*** (−16.90)	−0.0009 (−0.89)	−0.005*** (−4.14)	−0.001 (−0.54)	0.005 (1.81)
Following (021Ua)			−0.010*** (−10.70)	−0.013*** (−10.79)	−0.013*** (−10.90)	−0.019*** (−7.65)
030Ta				0.017** (3.27)	0.018*** (3.53)	0.020*** (3.93)
120Db				0.035*** (6.05)	0.035*** (−6.01)	0.031*** (5.35)
111Ub					−0.003** (−3.04)	−0.003** (−3.04)
111Ua					−0.004* (−2.41)	−0.01*** (−3.95)
111Da						0.009*** (2.94)
N	115,986	115,986	115,986	115,986	115,986	115,986
Pseudo r^2	0.034	0.062	0.069	0.083	0.084	0.085

Beta coefficients are reported with *z*-statistics in parentheses.
* $p < 0.05$.
** $p < 0.01$.
*** $p < 0.001$.

retweet in Models 4–6 (both $p < 0.001$, beta (030Ta) = ∼0.02, beta(120Db) = ∼0.03). In Model 6, three of the 111 MAN configurations (111Ub, 111Ua, and 111Da) had a significant effect on propensity to retweet. The coefficients for each of these 111 MAN configurations reflect their interaction with several of the more basic configurations in this model (mutual, followers and following). The exact meaning of these 111 configurations will be interpreted in the discussion below.

6. Discussion

6.1. Popular individuals

Our first hypothesis (H1) was '*That popular individuals will retweet more.*' The purpose of this hypothesis was to test the relative merits of the 'efficient-hub' and the 'inefficient-hub' theories of highly connected individuals. We operationalized the concept of 'popular individuals' with three measures: followers, following, and total tweets.

The results supported a version of the 'inefficient-hub' conception of highly connected individuals: all three measures of 'popular individuals' showed negative correlations with propensity to retweet. The overwhelmingly dominant driver of this effect is the variable 'following' (021Ua): when following (021Ua) is included in Model Three, it renders followers (021Cb) non-significant. In addition, in Models 3–6, both 'tweets' and 'followers' have substantially less explanatory power than the 'following'. Overall, the central importance of 'following' suggests that the inefficiency of popular individuals at retweeting is caused by difficulties processing the large quantities of information they receive: hubs have a limited ability to listen, and there probability of retweeting any particularly message they receive substantially decreases when they are following a large number of Twitter users.

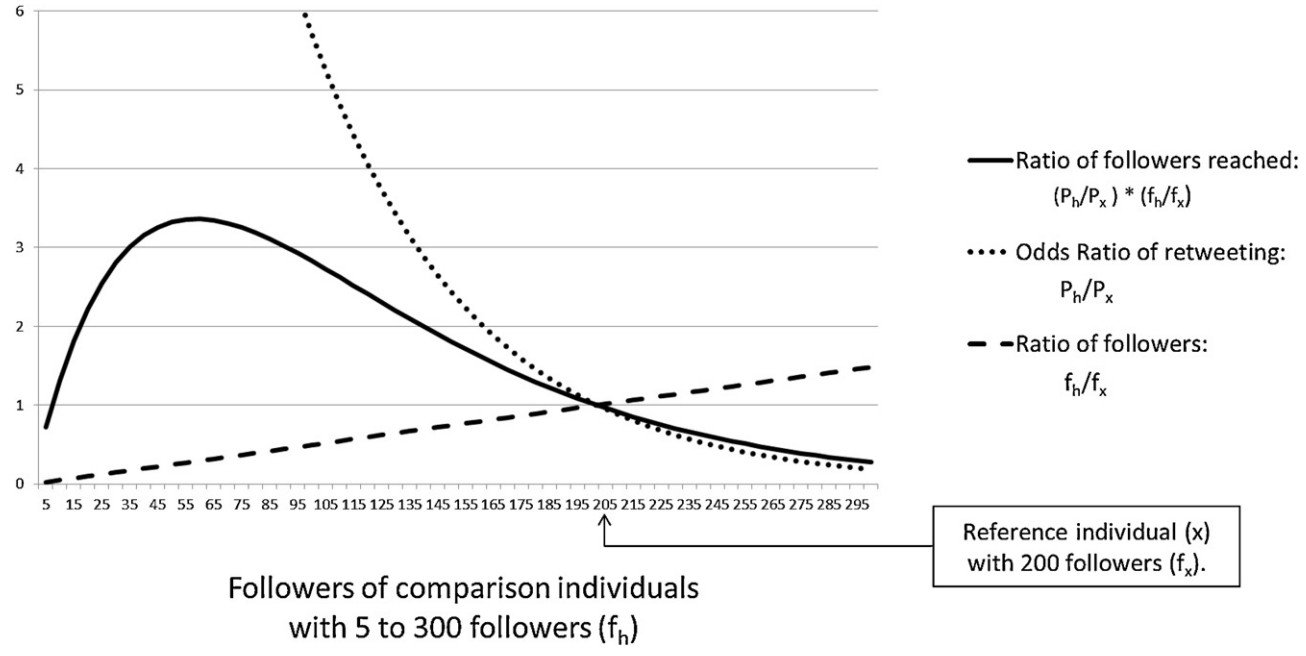

Fig. 2. The popularity trade-off: popularity increases the number of followers (f_h/f_x) but decreases the odds of retweeting (P_h/P_x). Followers reached ($(P_h/P_x) \times (f_h/f_x)$) peaks for individuals with just 60 followers. *Notes*: Calculations are done on the parameter estimates in Model 4. All variables are included. The reference individual's (x) covariates (graph counts) are equal to the mean plus two standard deviations (they simulate an 'influential' type person because their popularity (follower/following counts) is in the top ∼5% of the population). The comparison individual's (h) covariates are fixed and equal to the reference individual (x) except for covariates for followers and following. Followers of the comparison individual (h) vary as per the x-axis. The following value for the comparison individual (h) starts at the reference individual's (x) value. It varies with the followers of h, with a slope equal to: (standard deviation of mean of following)/(standard deviation of mean of followers). Other equations for the slope were tested and made no difference to the final results.

The effects of these popularity variables are substantial: The first 50 persons a user is following lowers a user's odds ratio of retweeting by 0.45: Users following one person retweet at twice the 'rate'[6] of users following 50 persons, and eight times the rate of users following 200 people. Similarly, the first 50 followers of a user lowers their odds ratio of retweeting by 0.60: Users with one follower retweet three messages for every two messages retweeted by people with 50 followers. Total tweeting has a similarly measurable effect: the first 500 messages sent during this five month period (total tweets) lowers a user's odds ratio of retweeting by 0.78.

The reader may question whether the decline in rate of retweeting really matters. The arguments outlined so far prove that there is a popularity trade-off: as an individual becomes more popular their rate of retweeting goes down. But this may be more than compensated for by their very increase in popularity. If we are interested in maximizing the number of people who receive a retweeted message we need to ask: Is a popular users' decline in propensity to retweet compensated for by their increased 'reach' to many followers? Fig. 2 addresses this question.

Fig. 2 examines this 'popularity trade-off', and tests whether a popular users decline in retweeting is compensated for by their increased 'reach' via their increased number of followers. The figure compares the effective 'reach' of a message sent to a reference individual (x), with the effective 'reach' of a message sent to a range of hypothetical individuals with between 5 and 300 followers.

Our reference individual (x) represents an 'influential', with covariates (graph counts) equal the population mean plus two standard deviations (placing them in the top 5% of individuals, thus simulating a 'popular individual'). Our comparison individuals (h) have the identical covariates (graph counts) to the reference individual (x) with exception of followers and following. Followers varies as with the x-axis and following varies with followers of h at a slope equal to the standard deviation of the mean number of following divided by the standard deviation of the mean number of followers (from Table 1).

Fig. 2 plots three graphs. Starting at the bottom of the legend, 'Ratio of followers' plots the ratio of the followers of h to the followers of x. This increases linearly and represents the relative increased or decreased number followers of the potential retweeters. 'Odds ratio of retweeting' is the ratio of the conditional probability of h retweeting divided by the conditional probability of x retweeting. Because of the negative coefficient (parameter estimate) for the number of followers and following, this odds ratio decreases with h. 'Ratio of followers reached' is the product of the values of the other two graphs. The graph plots the number (on average) of individuals receiving the message via individual h, for every one individual receiving the message via individual x.

What is Fig. 2's answer to our question? Does increased 'reach' compensate for decreased 'retweeting'? The answer is 'No'. 'Ratio of followers reached' is the key graph in Fig. 2, and it reaches a maximum when an individual has 60 followers. A person with just 60 followers reaches just under 3.5 times as many individuals with the messages they receive than does an individual (x) with 200 followers. This peak is almost exactly equal to the number of followers (60) of the average member of the population.

This pattern suggests that, first, the 'inefficient hub' effect is real and has a net negative effect: messages sent to 'popular' individuals actually reach less people. Second, to the extent that successful viral activity requires efficient viral retransmission, one would expect that a large part of the successful viral activity in the Twitter system is taking place between and through very 'ordinary' users (with just 60 or so followers).

[6] Where 'rate' refers to the number of retweets per messages received.

6.2. Community structures

We put forward three theories about the relationship between community structure and social contagion. The first theory emphasized the negative effect of community structures on contagion: it argued that messages passed within a community will tend to be redundant, therefore lack novelty, and will thus be associated with a lower rate of message retransmission. The second (similarity) and third (bonding) theories emphasized the positive effect of community structures on contagion: individuals within a community tend to be similar, and thus increasing the relevance of, and attention paid to, each other's messages; and individuals resending messages within communities tend to get much higher social bonding rewards from such messaging.

We proposed to test these theories of community structure and social contagion with our second and third hypotheses (H2 and H3). These were concerned with the effects of reciprocal ties and triadic structures on the propensity of receivers to retweet.

We operationalized hypothesis 2 (mutual ties) with two main measures: mutual tie (102a) and count of mutual friends (111Dc). We operationalized hypothesis 3 (triadic ties) through a variety of 'types' of triadic ties: '030Ca' (the 'egalitarian group'), '030Tb' (the 'group of disciples'); '030Ta' (the 'group of leaders'); '030Tc' (the 'hierarchical group'); and '210c' (the 'close community'). We also pointed out that both hypotheses 2 and 3 are partially operationalized in a large number of other subgraphs that incorporate the basic subgraphs (listed above) into a higher order subgraph (i.e. more complex subgraph).

At the general level, the results show a positive relationship between community structures and social contagion. These effects are particularly pronounced for the graph statistic 'mutual ties' (102a). A mutual tie between the sender and potential retweeter shows a strong positive correlation with the propensity to retweet a message: a receiver with a mutual relationship with the sender (origin of message) will retweet three messages for every one message retweeted by a receiver without a mutual tie to the sender. Configuration 111Dc (hypothesis 2b) did not show a significant relationship with retweeting.

The effects of triadic structures are more complex to disentangle. Table 2 shows that there is a general positive relationship between triadic structures and retweeting. Table 3 show that when all subgraphs are included in the model, two particular triadic structures are statistically significant: 030Ta ('group of leaders' or 'shared followers') and 120Db.

The increase in retweeting associated with configuration 030Ta ($p < 0.001$, beta = 0.02) is substantial: The first 50 shared followers increase the odds of retweeting by 2.7: People with 50 shared followers (030Ta) retweet approximately three times the 'rate' of people with only one shared follower. People with 100 shared followers (030Ta) retweet at 7.2 times the rate of people with only one shared follower. The increase in retweeting associated with configuration 120Db ($p < 0.001$, beta = 0.03) is also substantial: The first fifty 120Db configurations an individual is embedded in increases the odds of retweeting by 4.3.

But what is the meaning of the positive association of 030Ta and 120Db with retweeting? We think there are five main clues that help us provide a meaningful interpretation of the parameter estimates for 030Ta and 120Db. The first clue is that the two configurations are almost identical – they differ by just one tie (the K to J directed tie on 120Db) – suggesting that they represent similar sociological effects. The second clue is that their parameter estimates/coefficients are in the same direction (both positive), suggesting that we are not witnessing interaction between these two configurations. The third clue is that both configurations have a structure where the original sender (I) is not necessarily a member of the 'community' of J and K: J and K are 'listeners' of the

original sender. J and K appear to be an 'audience' for I's tweets, but I does not follow either J or K in either configuration 030Ta or 120Db. The fourth clue is that the original sender (I) and the potential retweeter (J) appear to be sending messages that appeal to a common audience: in both configuration 030Ta and 120Db, the shared third partner (K) is a follower of both J and I. The fifth and final clue is that 120Db places the emphasis on the community (the mutual tie) of successful retweeters is likely to be between J and K (not with I).

These characteristics of 030Ta and 120Db reinforces the conception that the best model for thinking about retweeting is one where retweeting is done with the explicit purpose of providing information to friends who are quite likely to have already received such information (hence, the information is "old news"). Retweeters are not put off by the potential redundancy of such information, and, in fact, are seemingly motivated (directly or indirectly) by the significant number of followers who share their status as followers of the postings of the original sender (I).

In relation to our theories, we find that community structures show a clear positive association with retweeting. At the dyadic level, close relations (a mutual tie) between the original sender and the retweeter considerably increases retweeting. At the triadic level, the key drivers of social contagion within community structures appear to be, firstly, having third party followers who also follow the original sender (030Ta), combined with, secondly, close mutual ties between the retweeter and these third party followers (120Db). Redundancy and lack of novelty of messages appear to either not negatively affect message contagion, or these effects are swamped by the positive effects of community structure – such as those that arise from user similarity or social bonding – on social contagion.

This finding – that structural redundancy and lack of novelty increases message contagion – places it in contrast to research on message content (Macskassy and Michelson, 2011) – which finds that user dissimilarity (anti-homophily) and message novelty increases message contagion. The contrasting findings may be squared in a number of ways, but probably the most likely explanation is that both effects are present in Twitter networks: users are both more likely to retweet messages from close relations/community structures (the finding of this paper), and more likely to retweet messages with original content (the finding of Macskassy and Michelson). This is certainly possible given that neither of the two studies have measured the other's variables.

6.3. The interaction of community structures and popularity (MAN 111 configurations)

Model 6 in Table 3 contains three higher order star configurations: 111Ub ($p < 0.01$, beta = −0.003), 111Ua ($p < 0.001$, beta = −0.01), and 111 Da ($p < 0.001$, beta = 0.009). These configurations capture the interaction effect of a mutual tie (between the sender (I) and potential retweeter (J)) and the number of followers or following of sender or potential retweeter. The interpretation of these results will be kept short for two reasons: firstly, the total explanatory power of these three variables is low. The improvement of the pseudo r^2 from Model 4 to Model 5 is just 0.001, which is approximately an improvement in explanatory power of just 0.1%. The second reason why the interpretation of these three variables will be kept short is that what appears to be represented by one of these coefficients (111Da) is statistical artifact of the modeling process.

The coefficient for 111Ub (beta = −0.003) needs to be interpreted with the coefficient mutual tie (beta = 1.10). Together, the coefficients for 111Ub and 'mutual tie' mean that potential retweeters (J) with a mutual tie to the original sender (I) are negatively affected by each user following the sender. The effect is substantial: with a mutual tie each additional user following the sender (I) reduces of the log of the odds of retweeting by 0.003. The qualitative interpretation of this is straightforward: the 'close friends' (mutual ties) of senders with large numbers of followers pay considerably less attention to these sender's tweets. Another way to think about it is to say that the positive effects of a mutual tie between sender and potential retweeter are completely cancelled out if the sender has more than 365 followers. This is probably an extension of the general principle of 'inefficient hubs', with senders with large numbers of followers unable to maintain meaningful relationships with large numbers of friends, and thus the mutual ties of these individuals become meaningless.

The coefficient for 111Ua (beta = −0.004 in Model 5, beta = −0.01 in Model 6) needs to be interpreted with the coefficients (i) followers (beta = −0.001 (not significant) in Model 5, beta = 0.005 ($p < 0.10$) in Model 6) (ii) mutual tie, and (iii) also 111Da (in Model 6). Focusing on Model 5 the coefficients for 111Ua and 'followers' mean that potential retweeters (J) with a mutual tie to the original sender (I) are more negatively affected by each user following the potential retweeter (J). The effect is substantial: with a mutual tie each additional follower of the potential retweeter (J) reduces of the log of the odds of retweeting by −0.005, while without a mutual tie each additional follower of the potential retweeter (J) has no effect on the odds of retweeting. The qualitative interpretation of this seems to be similar to that of 111Ub: mutual ties loose their uniqueness and their importance when potential retweeters have a large number of followers. Again this appears to be an extension of the principle of 'inefficient hubs': potential retweeters with a large number of followers (i.e. a large audience or a large number of fans) find themselves unable to maintain meaningful relationships with their mutual ties, and thus the mutual ties have less effect.

The coefficient for 111Da (beta = 0.009) appears to be largely a statistical artifact. In modeling not reported here, 111Da is not significant in models where 111Ua is not present, and when the two coefficients are present, they basically balance each other with coefficients in opposite directions of similar magnitude (a telltale sign of a statistical artifact). In addition, because the number of followers and the number of following of each user is so highly correlated, these two configurations are very highly correlated: Pearson's $R = 0.91$. While we present 111Da in Table 3 Model 6 for the purposes of full disclosure, we do not think the current modeling of 111Da can be given a confident scientific interpretation.

7. Conclusion

The social structure of a community will affect the pattern of the spread of a contagious idea within that community. But what is the exact relationship between local social structures and the successful spread of contagious ideas and behaviors?

We examined the effect of local social structure on the retweeting behavior of a community of Twitter users over a five month period. We focused on two broad categories of local social structure: popular individuals (those with large numbers of friends and high volumes of communication) and community structures (which occur between individuals with reciprocated ties or ties to common third parties).

We operationalized the concepts of popular individuals and community structures by counting subgraphs – small patterns of ties and nodes – in the networks of senders and retweeters.

We found that popular individuals have a significantly lower likelihood of retweeting, particularly when they are following a large number of individuals. Individuals following more than sixty users show a serious decline in both their propensity to retweet any particular message they receive, and also in the final 'reach' (number of users subsequently receiving) of any messages they are sent.

We interpret this as evidence of a type of information overload, a limitation on users' listening capacity, and, thus, as general support for the theories of inefficient hubs.

We find that community structures, particularly reciprocal ties, substantially increase social contagion. Individuals with mutual (reciprocal) ties to the sender have a significantly higher likelihood of retweeting a message. Individuals embedded in certain triadic community structure (030Ta and 120Db) showed a strong propensity to retweet messages. We argue these results are partial evidence against theories that emphasis the negative effect of community structures and message redundancy (and therefore lack of message novelty) on social contagion: there is no significant negative effect on retweeting for any of the community structure parameters measured. Instead, the evidence is that community structures have a positive effect on social contagion, and that, in fact, users are more likely to retweet information that is redundant "old news", and which their followers have already heard. We speculate that social contagion is higher within communities because users in a community are both more similar (and therefore messages more relevant) and because users in a community are better able to socially bond through message retransmission (reinforcing identity, cohesion, prestige and so forth).

We also find an interaction between certain community structures – particular mutual ties – and certain types of popularity (111Ua and 111Ub). We find that both senders and potential retweeters with large numbers of followers (which we could think of as fans or an audience) to lose the positive effects of a mutual tie between the sender and potential retweeter: the larger the number of followers of the sender or potential retweeter, the less the positive effect of having a mutual tie. We argue this is further evidence of the 'inefficient hubs' theory, with users with large numbers of followers having much more 'token' mutual friendships, reflecting the reality of having only limited resources to invest in real friendships.

Conflict of interest

There are no known conflicts of interest.

References

Barabasi, A., 2002. Linked: The New Science of Networks. Perseus Publishing.

Barabasi, A., Albert, R., 1999. Emergence of scaling in random networks. Science 286 (October (15)), 509–512.

Burt, R.S., 1992. Structural Holes: The Social Structure of Competition. Harvard University Press, Cambridge, MA.

Burt, R.S., 2005. Brokerage and Closure: An Introduction to Social Capital. Oxford University Press, Oxford.

Cha, M., Haddadi, H., Benevenuto, F., Gummadi, K.P., 2010. Measuring user influence in twitter: the million follower fallacy. In: Proceedings of the Fourth International AAAI Conference on Weblogs and Socia Media.

Frank, O., Strauss, D., 1986. Markov graphs. Journal of the American Statistical Association 81, 832–842.

Freeman, L.C., 1979. Centrality in social networks: conceptual clarification. Social Networks 1, 215–239.

Gladwell, M., 2000. The Tipping Point: How Little Things Can Make a Big Difference. Little Brown, New York, NY.

Golub, B., Jackson, M.O., 2007. Naïve learning in social networks: convergence, influence and wisdom of crowds. Social Science Research Network Electronic Paper Collection: http://ssrn.com/abstract=994312.

Granovetter, M., 1973. The strength of weak ties. American Journal of Sociology 78, 1360–1380.

Granovetter, M., 1983. The strength of weak ties: a network theory revisited. Sociological Theory 1, 201–233.

Helleringer, S., Kohler, H.-P., 2005. Social networks, perceptions of risk, and changing attitudes towards, HIV/AIDS. New evidence from a longitudinal study using fixed-effects analysis. Population Studies 59 (3), 265–282.

Hill, S., Provost, F., Volinsky, C., 2006. Network-based marketing: identifying likely adopters via consumer networks. Statistical Science 21, 256–276.

Holland, P.W., Leinhardt, S., 1970. A method for detecting structure in sociometric data. American Journal of Sociology 70, 492–513.

Holland, P.W., Leinhardt, S., 1981. An exponential family of probability distributions for directed graphs (with discussion). Journal of the American Statistical Association 76, 33–65.

Hunter, D., 2007. Curved exponential family models for social networks. Social Networks 29, 216–230.

Johnson Brown, J., Reingen, P.H., 1987. Social ties and word-of-mouth referral behavior. The Journal of Consumer Research 14, 350–362.

Katz, E., Lazarfeld, P.F., 1955. Personal Influence; The Part Played by People in the Flow of Mass Communications. Free Press, Glencoe, IL.

Leskovec, J., Adamic, L.A., Huberman, B.A., 2007. The dynamics of viral marketing. ACM Transactions on the Web 1 (May (1)).

Macskassy, S.A., Michelson, M. Why do people retweet? Anti-homophily wins the day. Proceedings of ICWSM 2011.

McFadden, D., 1973. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), Frontiers in Econometrics. Academic Press, New York.

Merton, R.K., 1968. Patterns of influence: local and cosmopolitan influential. In: Merton, R.K. (Ed.), Social Theory and Social Structure. Free Press, New York, NY, pp. 441–474.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U., 2002. Network motifs: simple building blocks of complex networks. Science 298, 824–827.

Robins, G., Pattison, P., Wang, P., 2006. Closure, connectivity and degrees: new specifications for exponential random graph (p^*) models for directed social networks, MelNet Website. http://www.sna.unimelb.edu.au/publications/publications.html (17 August).

Robins, G.L., Pattison, P., Elliott, P., 2001a. Network models for social influence processes. Psychome 66, 161–190.

Robins, G.L., Elliott, P., Pattison, P., 2001b. Network models for social selection processes. Social Networks 23, 1–30.

Rosen, E., 2000. The Anatomy of Buzz: How to Create Word-of-Mouth Marketing. Doubleday, New York, NY.

Snijders, T.A.B., 2002. Markov chain Monte Carlo estimation of exponential random graph models. Journal of Social Structure 3 (2).

Snijders, T.A.B., Steglich, C.E.G., Schweinberger, M., 2007. Modeling the co-evolution of networks and behavior. In: Montfort, K.V., Oud, J., Satorra, A. (Eds.), Longitudinal Models in the Behavioral and Related Sciences. Routledge, pp. 41–71.

Snijders, T.A.B., van de Bunt, G.G., Steglich, C.E.G., 2010. Introduction to actor-based models for network dynamics. Social Networks 32, 44–60.

Snijders, T.A.B., Pattison, P., Robins, G.L., Handcock, M., 2006. New specifications for exponential random graph models. Sociological Methodology 36, 99–153.

Wasserman, S., Pattison, P., 1996. Logit models and logistic regressions for social networks, I. An introduction to Markov graphs and p^*. Psychometrika 61, 401–425.

Wasserman, S.S., 1975. Random directed graph distributions and the triad census in social networks. National Bureau of Economic Research (NBER) Working Paper Series, November.

Watts, D.J., Dodds, P.S., 2007. Influentials, networks and public option formation. Journal of Consumer Research 34, 441–458.

Watts, D.J., 2007. Challenging the influentials hypothesis. WOMMA Measuring Word of Mouth 3, 201–239.

White, H.C., Boorman, S.A., Breiger, R.L., 1976. Social structure from multiple networks. I. Blockmodels of roles and positions. The American Journal of Sociology, 81.

Wu, F., Huberman, B.A., 2007. Novelty and collective attention. PNAS 104 (November (45)), 17599–17601.

Yang, J., Leskovec, J., 2011. Patterns of Temporal Variation in Online Media. In: WSDM'11, February 9–12, Hong Kong, China.