

# Uncovering Social Network Structures through Penetration Data

Yaniv Dover, Jacob Goldenberg,

The Jerusalem School of Business Administration, Hebrew University, Jerusalem, Israel 91905  
[yanivd@phys.huji.ac.il, msgloden@huji.ac.il]

Daniel Shapira

The Guilford Glazer School of Management, Ben-Gurion University, Beer Sheva, Israel, 84105  
[shapirad@bgu.ac.il]

## Abstract

We propose a method for uncovering the structure of the adopters' network underlying the diffusion process, based on penetration data alone. By uncovering the traces that this network leaves on the dissemination process, the degree distribution of the network can be estimated. We show that the network's degree distribution has a significant effect on the contagion properties.

Ignoring the network structure introduces significant errors to estimated diffusion parameters and may lead to flawed assessments of the magnitude of the contagion process. In three studies we validate the proposed method using data for known mapped networks and the adoption process propagating on them.

## Acknowledgments

We would like to thank Barak Libai, Yoram Louzon, Lev Muchnik, Oded Netzer, Arvind Rangaswamy, Sorin Solomon, and Olivier Toubia for their constructive comments and suggestions. This research was supported by the Israel Science Foundation, The Kmart International Center of Marketing and Retailing; The Davidson Center, Hebrew University of Jerusalem; The Horowitz Association; and the Center for Complexity Science.

# 1. Introduction

It is accepted that the structure of the network has a major impact on the diffusion process (Van den Bulte and Wuyts 2007, Shaikh et al. 2006, Mayzlin 2002, Newman et al. 2006, Katona et al 2009, Katona and Sarvary 2008, Hill et al. 2006, Goldenberg et al. 2009b). However, the structure of the network is invisible in most cases, and rough assumptions of the structure (e.g., homogeneity of the market, separation to two markets, etc.) are often used when developing a diffusion model. Such strong assumptions ignore the fact that network structure has been shown to affect several important aspects of product penetration; for example, Krackardt (1996) theoretically demonstrated how network structure can affect adoption patterns and how marketers can exploit this to optimize penetration. Similarly, most diffusion models to date have not incorporated the degree distribution of the network (the number of social ties of each individual).

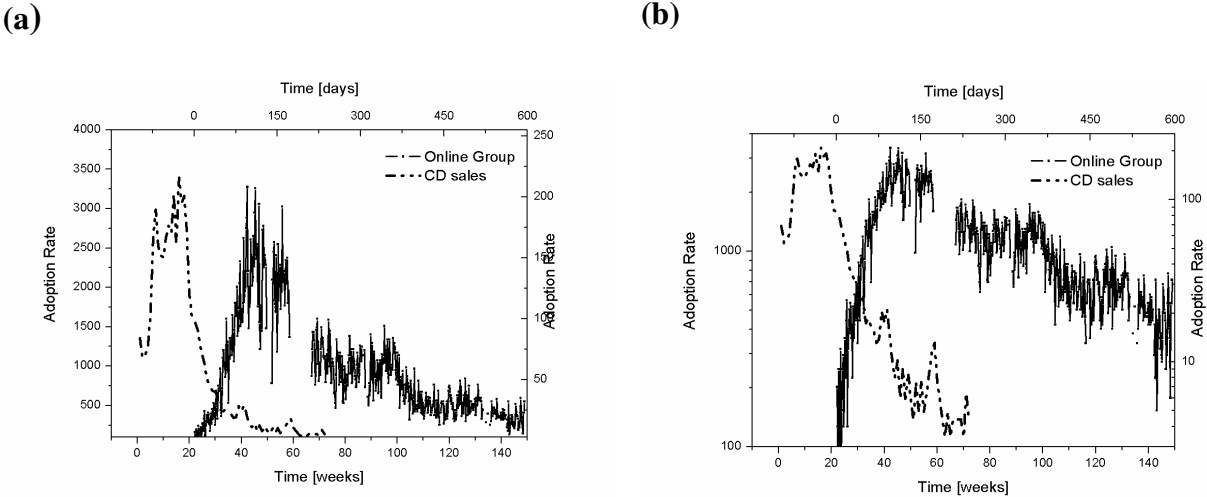
Since the magnitude and the speed of the contagion process depend strongly on the degree distribution of the underlying network (as we show in section 3.1), it is important for marketers to have knowledge of the distribution at hand so its role can be properly isolated from other factors affecting contagion. We show that the contagion depends linearly on the average number of neighbors and the standard deviation of the degree distribution. Thus, the presence of heavy tails in the distribution (e.g., the common scale-free distribution) can introduce errors of several orders of magnitude in the estimation of diffusion coefficients. Resulting erroneous managerial beliefs, such as over- or under-estimations of contagion force (if based on penetration data), can lead to erroneous marketing actions (e.g., investing in buzz programs or advertising when the opposite action is needed). Furthermore, the degree distribution contains complete information on the existence, the number, and the degree of heavily linked consumers, (e.g. social hubs), which have been shown to significantly affect the diffusion process (see e.g.,

Goldenberg et al. 2009b, Katona et al. 2009; note a different view on the limited nature of this influence, Watts and Dodds, 2007).

The purpose of this paper is to propose a method for revealing the properties of the network underlying the diffusion process (i.e., the type of degree distribution such as scale-free, normal, uniform etc.), as well as the two first moments of its degree distribution, based solely on penetration data. This information can then be incorporated in the growth model, and used to more precisely estimate the forces that drive penetration.

An example of the influence of network structure on the pattern of dissemination is illustrated in Figure 1. Consider two different diffusion processes: (a) Weekly sales of a certain music CD (the Dink album; taken from Moe and Fader 2001) and (b) The adoption rate of an online social network group (daily counts).

**Figure 1 Diffusion Patterns of CD sales (dashed line) and online group adoption (unbroken line) on a) normal and b) log scale.**



Plotting the diffusion curves on log-linear scale (Figure 1b) emphasizes the differences between the patterns (e.g. in the asymmetry and slopes of the curves), in comparison to the linear case (Figure 1a). This results from the exponential nature of the diffusion process.<sup>1</sup> In this

<sup>1</sup> Explanations of the need to use log scale are elaborated further below.

paper we demonstrate that the differences between these two adoption curves stem from the traces of the underlying network structures that are imprinted on and cause distinctions between the respective penetration processes. Analysis of these traces discloses estimates of the key properties of these networks' degree distributions, and sheds light on how they influence penetration, and how identification of the degree distribution allows us to correct the contagion estimations. Consider a theoretical case in which two products possess the same intrinsic contagion potential, i.e. once exposed to the product, the average consumer has, more or less, the same probability of adopting it through an interaction with another adopter (assuming all other adoption considerations and factors are equal). However, if Product A diffuses over a scale-free network, while Product B diffuses over a different, for example Poissonian network (for a detailed explanation, see section 3.1), the measured growth rate of Product A will be dramatically higher than that of Product B. A firm may, then, deduce that Product A has better “viral strength” in terms of contagion and will ultimately be more successful than Product B, which would be a costly misinterpretation.

Specifically for the penetration processes depicted in Figure 1 above, we show that the penetration curves, imprinted with network traces, suggest that the online group network has a scale-free structure while the CD sales fit a Gaussian-like degree distribution network structure. This difference implies that, unlike the case of the CD sales, social hubs play an active role in the propagation of the online group membership, and that the estimated diffusion internal coefficient is dominated by the variance in the number of social ties in the underlying network (see section 3.1). As in the above case, we find that even for products of a similar nature, the active social network (the subset of individual who actually adopted) underlying the respective diffusion processes may be entirely different than the structure of the potential market network, which is often assumed to be scale free.

The remainder of this paper is as follows: In section 2 we review past research and the background related to our work. In section 3 we develop the analytical baseline for our work and introduce our method for network reconstruction using aggregate penetration data. In section 4 we apply the network reconstruction procedure to simulated data, and in section 5 further test the reconstruction method through a variety of empirical cases. In section 6 we put the method to a more stringent test and use real network data to validate the uniqueness of the uncovered network. We offer our conclusions in section 7.

## **2. Background**

Traditionally, most diffusion models did not incorporate the structure of the consumers' network. In recent years, perhaps due to increasing availability of empirical data, such models have emerged to answer the question of how network structure affects new product diffusion. In general, social network research can be classified into studies of network formation (Allatta et al. 2009, Alon 2007, Barabasi 2003, Katona and Sarvary 2008, Kossinets and Watts 2006, Newman 2003) and studies that examine how network structure influences dissemination processes (Goldenberg et al. 2001, Goldenberg et al. 2009b, Katona et al. 2009; Shaikh et al. 2006, Van den Bulte and Joshi 2007, Van den Bulte and Wuyts 2007).

In response to the call in the literature for deeper investigations of what emerged as increasingly detailed structures of interpersonal connections within the consumer network (e.g., Mahajan et al. 1990, Mahajan et al. 1993), recent work suggests the importance of knowledge of the network structure for marketers. Network structure has been shown to have significant effects on the flow of information and influence (Katona and Sarvary 2009, Mayzlin and Yoganarasimhan 2009, Stephen and Toubia 2009). For example, one source of network impacts has been identified as a category of consumers known as the influentials (Katz and Lazarsfeld 1955, Van den Bulte and Joshi 2007) whose single base of influence is their extraordinary great

number of links to other individuals. While the actual influence of the great number of links of influentials has been questioned by Watts and Dodds (2007), more recent empirical evidence confirms that influentials accelerate the diffusion process (Goldenberg et al. 2009a).

It was shown that network effects have concrete economic value (Gupta 2006), i.e. the mere existence of network links adds monetary value per market. In fact, Stephen and Toubia (2009) have shown that in some cases specific network structures, such as those including consumers who have a single link each, might even depreciate economic value. In another application, Hill et al. (2006) have shown that knowledge of the consumers' network structure data can significantly improve the firm's capability of predicting consumers' likelihood of purchase. Finally, Shaikh et al. (2006) used a generalized diffusion model to evaluate the structure of small-world networks and found that structure has an important affect on the temporal aspects of new product diffusion; as a result, knowledge of network structure has a considerable affect on contagion parameter estimations. They also demonstrated that ignoring network effects may lead to incorrect interpretations of penetration data.

One common problem in the study of network structure is that networks of social influence are usually invisible or extremely hard to map (Rangaswamy et al. 2007). Despite this inherent opaqueness, marketers and researchers typically assume that dissemination processes propagate over the entire overt network. This assumption, however, is questionable, even based solely on the intuition that no innovation is adopted by all members of a social network (even if all members of the network are exposed to it).

We argue here that, for any given social network, diffusion frequently involves only a subset of the overt network, and the degree distribution of this subset may be different than the degree distribution of the network as a whole (similar to Stumpf et al. 2005). For example, a diffusion process can spread on a scale-free network but since the process spreads only on a subset of nodes, this active network may possess a Gaussian network structure. Indeed, it was

found that internet chain letters propagate in a “narrow but very deep tree-like pattern continuing for several hundred steps” rather than fanning out widely, reaching many people in a very few number of steps as expected by small-world principles (Liben-Nowell and Kleinberg 2008).

An important implication of knowing the real (active) network structure concerns accuracy in understanding and estimating the penetration process, which is subjected to biases even in the absence of oversimplified network assumptions (see Van den Bulte and Lillien 1997, 2001). When the magnitude of contagion is estimated, using a penetration pattern, the effect of network structure is generally not taken into account.

Recently the reconstruction of unknown network properties from other known properties (e.g. number of links, links probability, etc.) has been attempted using maximum-likelihood methods (Garlaschelli and Loffredo 2008, Ramasco and Mungan 2008) or through a combination of aggregate data and Bayesian model selection techniques (Trusov and Rand 2009). On the micro level, Braun and Bonfrer (2009) have developed a method to uncover the hidden dyad-level interdependence between consumers. In this paper we present a different approach for uncovering unknown properties of the network. Rather than testing consistency with other known topological properties of the network as a reference point, we use penetration data to reconstruct the underlying network.

### **3. Methodology**

#### **3.1 The analytical baseline**

Consider the simplest case of social influence, in which individuals affect each other equally, the market is homogenous for both external and internal influences, and both effects are constant in time. These assumptions lead to the following differential equation (Bass 1969):

$$\frac{dN(t)}{dt} = \left( P + Q \cdot \frac{N(t)}{M} \right) \cdot (M - N(t)) = MP + (Q - P)N(t) - \frac{Q}{M} N(t)^2. \quad (1)$$

Here,  $N(t)$ ,  $M$ ,  $P$ ,  $Q$ , are the cumulative adopters at time  $t$ , the total population, the external force coefficient, and the internal force coefficient, respectively.

For the early stages of the process (i.e. for low values of  $t$ ), the cumulative number of actual adopters is relatively small in comparison to the market potential  $M$ , so that  $N(t) \ll M$ , allowing the following first-order linear approximation:

$$\frac{dN(t)}{dt} = MP + (Q - P)N(t) - Q \left( \frac{N(t)}{M} \right) N(t) \approx MP + (Q - P)N(t). \quad (2)$$

Specifically, at early stages of the process, the adoption rate demonstrates exponential growth of the following form:

$$\frac{dN(t)}{dt} \propto e^{(Q-P)t}. \quad (3)$$

For much later stages (high values of  $t$ ),  $N(t) \sim M$  and hence  $\delta(t) = M - N(t) \ll M$ , allowing for another linear approximation:

$$\frac{dN(t)}{dt} = \left( P + Q \left( 1 - \frac{\delta(t)}{M} \right) \right) \delta(t) \approx (P + Q) \delta(t) = (P + Q)(M - N(t)). \quad (4)$$

Thus, towards the end of the process, the adoption rate declines exponentially:

$$\frac{dN(t)}{dt} \propto e^{-(Q+P)t}. \quad (5)$$

Both slopes are exponential and therefore become linear on a log scale (this fact will be used below).

It is common to assume that the internal force is larger than the external force ( $P \ll Q$ ) (Farley et al. 1995), the temporal dynamics for a fully connected market (where a network is not yet assumed) is approximately symmetrical around the peak, and the absolute value of the exponent is approximately the same for both growth and decline. When network effects are taken into account, this temporal symmetry collapses, and the deviations from symmetry provide initial indications of the network structure.

Consider a random network in which the nodes are connected randomly, with an equal probability for any pair of nodes to be linked, up to the limit of  $k$  neighbors (degrees) for each node (Erdős 1959). The parameter  $k$ , a node's number of neighbors or network degree, is retrieved from the network degree distribution  $P_k$ . At each time step  $\Delta t$ , any consumer who has adopted the product in question has a probability  $q\Delta t$  of influencing her neighbors to adopt. In addition, each node has a probability of  $p\Delta t$  per time step to be influenced by the external force (marketing forces) to adopt the product. We also assume that both internal and external influence rates are completely homogenous in time and space and that the network is undirected (influence is bi-directional). Generally, in the event that a potential adopter has  $x$  neighbors who have already adopted the product, her probability of adopting the product in time interval  $\Delta t$  is  $(p + xq)\Delta t$ . Therefore, the dynamics of the expected rate of adoption over a network in the continuous limit (i.e. where the time step duration  $\Delta t$  approaches zero), is given by:

$$\frac{dN(t)}{dt} = \sum_x H_x(t) \cdot (p + xq). \quad (6)$$

Here,  $H_x(t)$  denotes the number of potential adopters of order  $x$  who are also neighbors to  $x$  adopters of the product at the time  $t$ . Naturally, the *total* number of potential adopters at a given time is:

$$\sum_x H_x(t) = M - N(t). \quad (7)$$

Formulated in this manner, the Bass equation can be viewed as a special case of Equation 6. Assuming that all individuals are connected to all other individuals, the order of all potential adopters at time  $t$  is  $x = N(t)$ , where  $H_x(t) = M - N(t)$ . Thus

$\frac{dN(t)}{dt} = (M - N(t))(P + \frac{Q}{M}N(t))$ , where  $P = p$  and  $Q = Mq$ . Namely, the slope of the growth in the Bass model (see Equation 5) is determined by the product of the individual-level contagion coefficient  $q$  and the entire market potential  $M$ .

In Appendix A we formally describe the model of diffusion on a generalized random network, assuming that the maximal degree in the network is significantly smaller than the size of the network (which is a realistic assumption for social networks). We calculate (in Appendix A) the functions  $H_x(t)$  together with a set of conditional probabilities of the form  $f_{k|x}(t)$  (where  $f_{k|x}(t)$  is the conditional probability that potential adopter of order  $x$  at time  $t$  has network degree  $k$ ). The conditional probabilities  $f_{k|x}(t)$  and the functions  $H_x(t)$  become the solutions of a closed system of coupled Ordinary Differential Equations where the initial conditions are determined solely by network size (i.e. market potential  $M$ ) and network degree distribution  $P_k$ . The analysis leads to expressions describing the adoption rates in two regimes: the growth regime in early stages of the diffusion process, and the decline regime in more advanced stages toward the conclusion of the process.

### **The growth stage.**

In the early stages of the process, because network architecture is random, the probability that two neighbors of a specific node are neighbors themselves is negligible (for networks that are sufficiently large compared to the maximal network degree). Hence, in the initial stages of the penetration process, there is no more than a negligible probability that any potential adopter has more than one neighbor who has already adopted the product. To calculate the exponential growth slope, it is possible to drop all high-order terms that include  $H_x(t)$  for  $x \geq 2$  and linearize the dynamic equations of diffusion on a generalized random network. Assuming that  $\tilde{Q} \gg p$ , we find (see Appendix A) that the adoption rate in the early stages of the process takes the following form:

$$\frac{dN(t)}{dt} = (M - N(t))p + qH_1(t) \approx Mp\left(1 - \frac{k_{avg}q}{\tilde{Q} - 2p}\right)e^{-pt} + Mp\frac{k_{avg}q}{\tilde{Q} - 2p}e^{(\tilde{Q}-3p)t}. \quad (8)$$

Here,  $\tilde{Q}$  is determined by the ratio of the second moment of the network degree distribution to the first moment, giving:

$$\tilde{Q} = q \left\{ \frac{\sum_{k=k_{min}}^{k_{max}} k^2 P_k}{\sum_{k=k_{min}}^{k_{max}} k P_k} - 2 \right\} = q \left( k_{avg} + \frac{\sigma^2}{k_{avg}} - 2 \right), \quad (9)$$

where  $k_{avg} = \sum_{k=k_{min}}^{k_{max}} k P_k$  and  $\sigma^2 = \sum_{k=k_{min}}^{k_{max}} (k - k_{avg})^2 P_k$  are the average and variance of the network degree distribution, respectively. Thus, where  $\tilde{Q} > 3p$  (e.g., Farley et al. 1995), the rate of adoption exhibits exponential growth of the following form:

$$\frac{dN(t)}{dt} \propto e^{(\tilde{Q}-3p)t}. \quad (10)$$

That is, for a given average, greater network degree variance leads to greater exponential growth. This occurs because highly connected individuals accelerate growth.

To offer some intuition on the rationale behind this growth expression, consider the simple case of  $p = 0$  where no external force exists. In this case, all adoptions occur among consumers exposed to internal influence (through neighbors who are adopters). As mentioned above, at the initial non-interactive stage, each exposed consumer has only one adopter neighbor, on average, who can influence her to adopt. Therefore, if an individual with network degree  $k$  adopts the innovation during the initial stages of the process, the number of exposed consumers is increased by  $k - 2$ , as she has  $k - 1$  neighbors who have not yet adopted the new product and she is also removed from the list of exposed consumers by becoming an adopter of the innovation. Taking into account the fact that nodes with a greater number of links have greater propensity, on average, to being exposed to a spreading adoption process, each new adoption

adds  $\sum_{k=k_{\min}}^{k_{\max}} (k - 2)\tilde{P}_k = \sum_{k=k_{\min}}^{k_{\max}} k\tilde{P}_k - 2$  individuals to the number of exposed consumers, where  $\tilde{P}_k$  is

the degree distribution among exposed consumers, given by the probability that a node will be a neighbor of another node with degree  $k$ . The distribution of the degrees of neighbors on random linked networks is known to be (Albert and Barabasi, 2002):

$$\tilde{P}_k = \frac{k \cdot P_k}{k_{avg}}, \quad (11)$$

where  $P_k$  is the degree distribution and  $k_{avg}$  is the average degree for the purpose of "normalization" (Note that this probability is not the original probability  $P_k$  that a certain node in the network will have degree  $k$ . The larger the degree  $k$  for a given node, the greater number of ties she has, and hence the probability that she is included in another individual's sample is skewed toward higher degrees). Hence, the total increase in the number of exposed consumers in a single time step is:

$$\frac{dH_1(t)}{dt} = \left( \sum_{k=k_{\min}}^{k_{\max}} k\tilde{P}_k - 2 \right) \cdot qH_1(t) = \left( k_{avg} + \frac{\sigma^2}{k_{avg}} - 2 \right) qH_1(t) \equiv \tilde{Q}H_1(t) \quad (12)$$

as a result of which, in the case of  $p = 0$ , the adoption rate becomes  $\frac{dN(t)}{dt} = qH_1(t) \propto e^{\tilde{Q}t}$ .

Adding external influence moderates the slope of the exponential growth, as indicated by Equation 8.

Equation 9 is the analytical measure of the dependence of estimated internal force on network structure: The coefficient depends linearly on the average number of neighbors, as well as on the standard deviation of the degree distribution. This explains why diffusion is accelerated when the network contains individuals with an exceptionally high degree. The presence of heavy-tailed distributions (i.e. existence of influentials) can change the average and standard deviation by several orders of magnitudes and consequently affect the imitation coefficient.

Thus, the same product diffusing on different networks can exhibit dramatically different  $\tilde{Q}$ s.

In order to assess the error caused by ignoring network structure, consider the most simple estimation of contagion size, using the Bass model. In this case, the effect of the internal force ( $Q$ ) is underestimated because the network structure multiplier  $k_{avg} + \frac{\sigma^2}{k_{avg}} - 2$  is replaced by

entire market potential  $M$ .

### **The decline stage.**

By the final stage of the adoption process (large  $t$ ), almost all individuals have adopted the innovation. The adoption map assumes the form of a network with isolated "holes," reflecting that most non-adopters are linked exclusively to adopters. This happens at the stage in which most of the network is occupied by adopters and thus, the order (number of friends that have already adopted) of almost all non-adopters becomes equal to their network degree (i.e. all the friends of nodes that have not adopted yet – have fully adopted). Moreover, the sole impact of product adoption by an individual with network degree  $k$  is a reduction in  $H_k(t)$ , the number

of potential adopters of order  $k$ . Hence, recalling that the probability per time step of potential adopter of order  $k$  to adopt is  $kq + p$ , we obtain:

$$\frac{dH_k(t)}{dt} = -(kq + p)H_k(t), \quad (13)$$

resulting in  $H_k(t) \propto e^{-(kq+p)t}$ . Therefore, the rate of adoption at the final stages of the process can be obtained from Equation 5 as follows:

$$\frac{dN(t)}{dt} = \sum_k H_k(t) \cdot (p + kq) \approx \sum_k C_k e^{-(kq+p)t}, \quad (14)$$

where  $C_k$  are time-independent coefficients. Namely, the rate of adoption is given by the sum of time-decaying exponentials, which is dominated by the slowest decaying exponent. That is, the adoption rate obeys the following decline:

$$\frac{dN(t)}{dt} \propto e^{-(k_{\min}q+p)t}, \quad (15)$$

where  $k_{\min}$  denotes the lowest degree in the network (The dynamics throughout the decline stage are derived directly from the general equations of diffusion on random networks in Appendix A).

### **The effect of network structure on the adoption curve.**

Recall that the resulting pattern of an adoption process that propagates on a random network exhibits exponential growth in the initial stages of the process, and exponential decline in the final stages of the diffusion. The growth and the decline slopes were shown to be, respectively:

$$x_1 = q \left( k_{avg} + \frac{\sigma^2}{k_{avg}} - 2 \right) - 3p \quad (16)$$

and

$$x_2 = -(k_{\min} \cdot q + p). \quad (17)$$

Here,  $k_{avg}$  and  $\sigma^2$  are the mean and the variance of network degrees, respectively, and  $k_{\min}$  is the lowest degree of the network. In cases where the network has a degree distribution with a

single scale, the low and the high degrees are of the same order, resulting in  $k_{avg} + \frac{\sigma^2}{k_{avg}} \approx k_{\min}$ .

Hence, considering the more common case when external influence  $p$  is small, the absolute values of the growth and decline slopes tend to be very close to each other and generate a relatively *symmetric* curve of adoption. In contrast, heavy-tailed degree distributions (e.g.,

scale-free distributions) that span over several scales of magnitude (where  $k_{avg} + \frac{\sigma^2}{k_{avg}} \gg k_{\min}$ )

should lead to an *asymmetric* curve of adoption. The slope of growth is much steeper while the slope of the decline exhibits a long-lasting temporal tail.

To illustrate the asymmetrical functional form of the penetration curve, we simulated diffusion on a 100,000-node-network for different classes of networks. Gaussian and uniform network diffusion processes are given in Figures 2a and 2b, respectively, to represent the category of narrow-network-degree distributions<sup>2</sup> (such as seen in Amaral et al. 2000). Another, more common category of degree distributions is the fat-tailed distribution (Newman 2005), which exhibits a non-negligible probability of hubs, also known as hyper-influentials (Goldenberg et al. 2009a, Watts and Dodds 2007), or consumers with an extraordinary number of friends. These fat-tailed distributions are commonly expressed in the form of scale-free (also known as power laws, see Barabasi and Albert 1999) and lognormal (Limpert 2001)

---

<sup>2</sup> Although the uniform network process does exhibit moderate asymmetry features.

distributions. Figures 2c and 2d demonstrate diffusion processes over lognormal<sup>3</sup> and scale-free networks, respectively.

As can be seen in Figure 2a-d, each network category imprints a unique pattern on the dissemination process, which is easy to identify on a semi-log scale (compared to the more commonly used linear scale). In cases where the network has a degree distribution with a single scale, such as Gaussian or Poissonian distributions, a minor tail of the adoption rate time series is created by the “lowest degree” nodes, which are the last to enroll in the process. The result is an exponentially decaying tail in the final stages of the dissemination process (see Figure 2a). In the case of a uniform distribution, the tail is more accentuated and longer-lasting in relation to the curve itself (Figure 2b), representing the relatively larger population of lowest degree nodes. The lognormal degree distribution has a longer tail, emanating from the peak itself, with several decaying exponentials due to the large lowest degree node population, and accelerated growth rate due to its heavy tail comprising highly connected nodes (Figure 2c).

Finally, since lowest degree nodes dominate the scale-free case, the tail is long, originates from the peak, and demonstrates a single exponential decay (distinctly log-straight, but in some rare cases demonstrates curviness) following sharp rapid growth (Figure 2d). These are unique patterns, as we demonstrate in our numerical and analytical studies below (see sections 4-6) and therefore are the latent fingerprints of the active network in each of these dissemination processes. This allows us to reconstruct the network degree distribution.

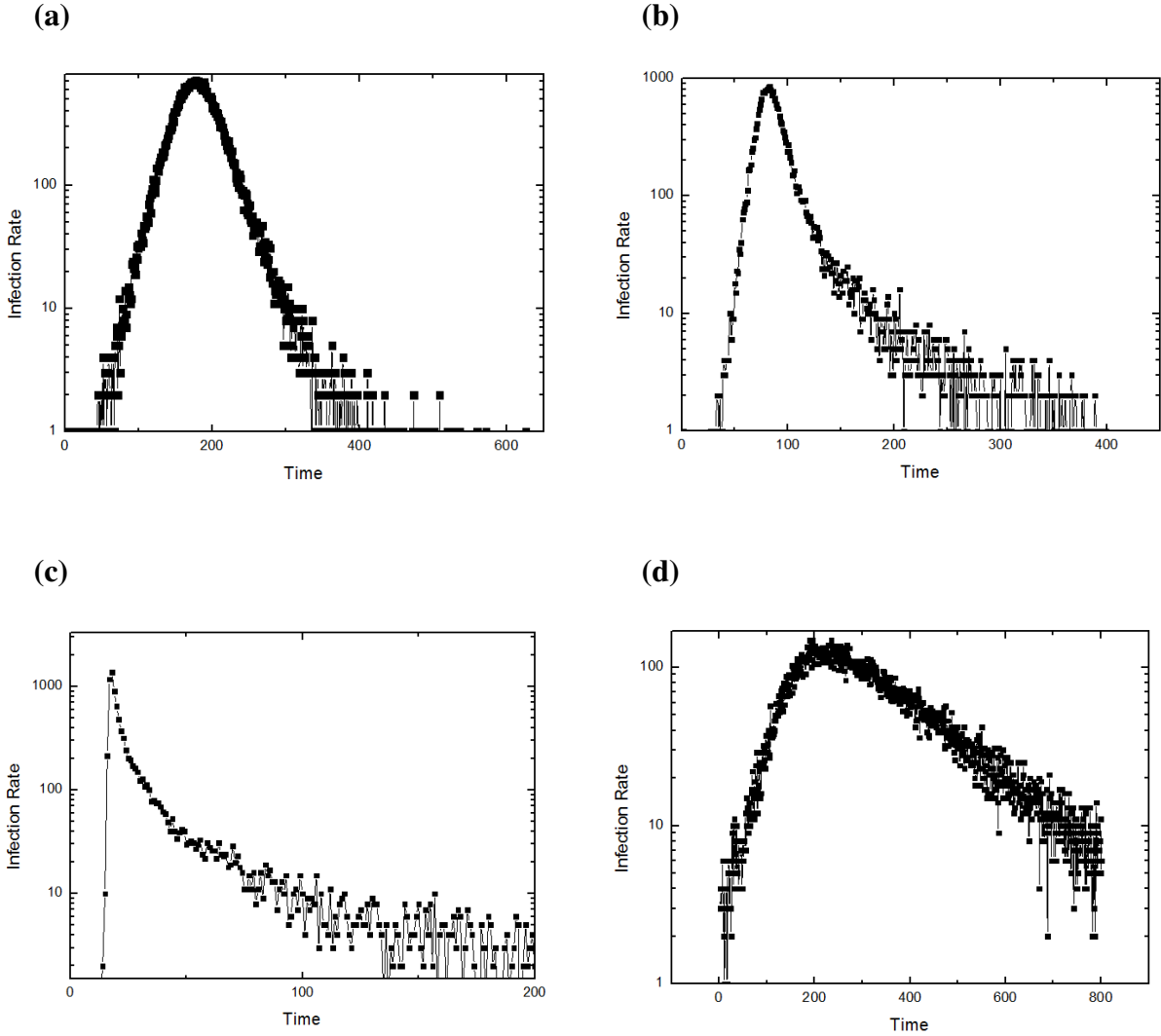
When network structure is not taken into account, efforts to measure diffusion by fitting diffusion-like models may produce erroneous results. For example, in light of equation 9, a marketer who customarily assumes that each consumer is connected to the *entire* market and not only to his local neighborhood might significantly underestimate the contagion coefficient.

---

<sup>3</sup> Lognormal networks have been identified in a variety of empirical contexts (Limpert et al 2001). These networks exhibit a "normal-degree distribution" over the **log scale** of degrees. Several explanations have been suggested for their existence (Limpert et al 2001 and references therein).

Moreover, fitting diffusion models without taking into account the asymmetry imposed by network structure may lead to inaccurate estimations.

**Figure 2 Diffusion Patterns for Different Network Types: a) Gaussian b) Uniform c) Log-normal d) Scale free.**



### 3.2 Estimating Network Degree Distribution

The procedure we employ to estimate the network degree distribution is based on the simplified assumption that diffusion occurs on an approximately random network of consumers. To model the adoption process, we use the parsimonious agent-based model (e.g., Garber et al. 2004), and consider four classes of networks: Gaussian/Poissonian, uniform,<sup>4</sup> lognormal, and scale-free networks. These distributions are representative of dissemination processes documented in the literature (Newman 2003; Newman et al. 2006; Amaral et al. 2000). In Table 1 we provide definitions of the distributions we use in the reconstruction method, including their parameters, and details regarding the estimations themselves. It should be noted that some networks cannot be perfectly associated with any "pure" category: Hybrid networks with degree distributions, for example, span more than one category (possibly, a Gaussian distribution for small degrees and a heavy scale-free tail of large degrees, although other hybrid combinations do exist). Nonetheless, we argue and also confirm empirically that of all candidate networks that could result in each diffusion pattern under investigation, the network type identified by our procedure has a degree distribution that is the closest to the degree of the active network underlying the diffusion process.

As indicated in Table 1, each network structure consists of two parameters (e.g. in the case of Gaussian degree distribution,  $\pi_1$  and  $\pi_2$  denote the average and the standard deviation, respectively). Thus, including the additional two parameters of the process (i.e. the external and internal forces,  $p$  and  $q$  respectively), the estimation involves a total of four parameters.<sup>5</sup>

The proposed procedure consists of a standard non-linear fitting process with modifications. Since the analytical solution is not in closed form (see Appendix A), we use simulations for the purpose of fitting. We iterate over simulation runs fitted to the empirical data, using maximum achievable goodness-of-fit measures as a criterion for reaching a solution. Since the non-linear,

---

<sup>4</sup> We also tested fit for the uniform-degree distribution, since it is a distribution that lies somewhere between light- and heavy-tailed.

<sup>5</sup> Since we perform the analysis on a given, full, diffusion curve we already know the market potential.

four-dimensional solution space may entail a costly, perhaps unfeasible search, we therefore use three analytical relations of the pattern itself: (1) the pre takeoff (2) the growth and (3) decline slopes. As a result, search is one-dimensional (or two dimensional, in the worst-case scenario of a scale-free network in which the growth analytical relation yields a tangled constraint). Next we describe the procedure to extract the three relations and the remainder of the estimation process.

Conditional on the existence of a pre-takeoff period, we can estimate external influence  $p$ . In the pre-takeoff stage, a very early stage of the process, the rate of adoption is approximately  $\frac{dN(t)}{dt} \sim \frac{dN}{dt} \Big|_{t=0} = Mp$ . Thus, since it is possible to estimate the average adoption rate during the pre-takeoff stage, which we assume to be  $Mp$  (the external force process is dominant in this stage), and the total cumulative number of adopters at the end of the process  $M$  is a known parameter, the external force  $p$  can be extracted as well.

In the previous sub-section we demonstrated that any diffusion process that propagates on a random network involves growth and decline stages resulting in two constraints due to their slopes  $x_1$  and  $x_2$  as indicated by Equations (14) and (15). These slopes are extracted from the aggregate-level adoption curve as described in Appendix B. We then separate the external influence from the network structure by retrieving empirical measures of the adjusted growth and decline slopes, and revise equations (14) and (15) to define the following system of two equations to solve for the three unknown parameters  $\pi_1$ ,  $\pi_2$ , and  $q$ :

$$z_1 = q \left( k_{avg} + \frac{\sigma^2}{k_{avg}} - 2 \right) \quad (18)$$

and:

$$z_2 = qk_{min} \quad (19)$$

Here,  $z_1 = x_1 + 3p$  and  $z_2 = -x_2 + p$  are the adjusted slopes of growth and decline, respectively (satisfying  $z_1 > z_2 > 0$ ), and  $k_{avg} = k_{avg}(\pi_1, \pi_2)$ ,  $\sigma^2 = \sigma^2(\pi_1, \pi_2)$  and  $k_{min} = k_{min}(\pi_1, \pi_2)$  are functions of the two parameters of the network degree distribution. In Table 1 we provide the explicit form of these functions for each network class. Equations (18) and (19) allow us to express two unknown parameters as a function of the third unknown parameter for all categories (with the exception of the scale-free network class, where only one unknown parameter can be analytically solved using the remaining two unknown parameters).

In the final stage of the procedure, as we indicated above, we use a computer simulation as a fitting function, to fit for the final parameter, subject to constraints (18) and (19). We used an ABM (agent-based models, Bonabeau 2002) to simulate a new product penetration process over a given network (Goldenberg et al. 2001). We then use the adoption pattern generated by the simulation as a fitting function for the data. Along with equations that describe parts of the pattern (see Equations (16) and (17)), we narrow the fit to a fit of one or two parameters, and thus significantly improve convergence accuracy. This estimation procedure is conducted for each network structure class where the reconstructed active network is defined as the parameter set that produces the highest goodness-of-fit score (in terms of R-squared measures) of the simulated network to the empirical adoption data.

**Table 1: Classes of network structures**

Network Degree Distribution	<u>Gaussian/Poissonian Networks</u>
	$P_k = \frac{1}{S\sqrt{2\pi}} e^{-\frac{(k-\mu)^2}{2S^2}}$
	<p>To guarantee that all network degrees are positive, we assume that <math>\mu - 3S &gt; 0</math>. (See comment below for cases where this condition is violated).</p> <p>The case of <math>\mu - 3S &lt; 0</math>: In this case, the form of the network degree distribution should be</p>

Poissonian ( $P_k = \frac{\mu^k e^{-\mu}}{k!}$ ) where the degree average and variance are given by  $k_{avg} = \mu$  and  $\sigma^2 = \mu$  respectively. Hence the growth constraint takes the form  $z_1 = 2q(\mu - 1)$  and for any choice of  $q$  one obtains  $\mu = \frac{z_1}{2q} + 1$ . The Poissonian degree distribution can be considered an anomaly of the Gaussian degree distribution in the case of small numbers. As for  $\mu > 10$  a Gaussian distribution with the parameters  $\pi_1 = \mu$  and  $\pi_2 = S = \sqrt{\mu}$  becomes a good approximation of the Poissonian distribution (subject to the continuity correction (Devore 1995)).

**Distribution parameters**

$\pi_1 = \mu$  : The network degree average.  $\pi_2 = S$  : The network degree standard deviation.

**Functions of the distribution parameters Constraints**

$$k_{avg}(\mu, S) = \mu, \quad \sigma^2(\mu, S) = S^2, \quad k_{min}(\mu, S) = \mu - 2S$$

The effective minimum number of neighbors  $k_{min}$  is the  $k$  that is approximately 2 standard deviations below the average.

The growth constraint:  $z_1 = q\left(\mu + \frac{S^2}{\mu} - 2\right)$ , The decline constraint:  $z_2 = q(\mu - 2S)$

**Chosen range of parameter values in the estimation procedure**

$q$  is chosen in the range between 0.0005 and 0.5.

$S = C(1 + \Delta)$  where  $C = \frac{1}{5q}(z_1 - 2z_2 + 2q)$  and  $\Delta = \sqrt{1 + \frac{5z_2(z_1 - z_2 + 2q)}{(z_1 - 2z_2 + 2q)^2}}$  while satisfying

the conditions  $C > 0$  and  $\Delta > 1$ .  $\mu = \frac{z_2}{q} + 2S$

**Uniform Networks**

**Network Degree Distribution**

$$P_k = \frac{1}{b-a} \text{ where } a \leq k \leq b.$$

**Distribution parameters**

$\pi_1 = a$  : The minimal degree of the network.  $\pi_2 = b$  : The maximal degree of the network.

**Functions of the distribution parameters Constraints**

$$k_{avg}(a, b) = \frac{a+b}{2}, \quad \sigma^2(a, b) = \frac{(b-a+1)^2 - 1}{12}, \quad k_{min}(a, b) = a.$$

The Growth constraint:  $z_1 = q\left(\frac{a+b}{2} + \frac{(b-a+1)^2 - 1}{6(a+b)} - 2\right)$ . The Decline constraint:  $z_2 = qa$

**Chosen range of parameter values in the estimation procedure**

$q$  is chosen in the range between 0.0005 and 0.5.

$$a = \frac{z_2}{q}$$

$b = C(1 + \Delta)$  where  $C = \frac{1}{8q}(6z_1 - 4z_2 + 10q)$  and  $\Delta = \sqrt{1 + \frac{16z_2(6z_1 - 4z_2 + 14q)}{(6z_1 - 4z_2 + 10q)^2}}$

while satisfying the conditions  $C > 0$  and  $\Delta > 1$ .

**Lognormal Networks**

**Network Degree Distribution**

$$P_k = \frac{1}{kS\sqrt{2\pi}} e^{-\frac{(\ln(k)-\mu)^2}{2S^2}}$$

**Distribution parameters**

$\pi_1 = \mu$  : The average of network degree logarithms.  $\pi_2 = S$  : The standard deviation of network degree logarithms.

**Functions of the**

$$k_{avg}(\mu, S) = e^{\mu + \frac{S^2}{2}}, \quad \sigma^2(\mu, S) = (e^{S^2} - 1)e^{2\mu + S^2}, \quad k_{min}(\mu, S) = e^{\mu - 2S}$$

**distribution parameters** The network degree logarithms are normally distributed with mean  $\mu$  and standard deviation  $S$ . Hence the effective minimum number of neighbors  $k_{\min}$  is approximately given by  $\ln(k_{\min}) \approx \mu - 2S$  (so that  $\ln(k_{\min})$  is about 2 standard deviations below the average of the network degree logarithms.)

**Constraints** The Growth constraint:  $z_1 = q \left( e^{\frac{\mu + 3S^2}{2}} - 2 \right)$ . The Decline constraint:  $z_2 = qe^{\mu - 2S}$

**Chosen range of parameter values in the estimation procedure**  $q$  is chosen in the range between 0.0005 and 0.5.

$$S = \frac{2}{3}(1 + \Delta) \text{ where } \Delta = \sqrt{1 + \frac{3}{2} \ln \left( \frac{z_1 + 2q}{z_2} \right)} \text{ while satisfying the condition } \Delta > 1.$$

$$\mu = \ln \left( \frac{z_2}{q} \right) + 2S$$

**Network Degree Distribution**

### Scale-free Networks

$$P_k = \frac{k^{-\alpha}}{Z}$$

Here  $a \leq k \leq b$  such that  $b \gg a$  and  $Z = \sum_{k=a}^b k^{-\alpha} \sim \int_a^b k^{-\alpha} dk = \frac{a^{-\alpha}}{1-\alpha}$ .

**Distribution parameters**

$\pi_1 = \alpha$  : The power exponent of the distribution, which is  $2 < \alpha < 3$  (Newman et al. 2006).

$\pi_2 = a$  : The minimal degree of the network.

The maximal degree of the network  $b$  defines the upper cut-off point of the distribution, so the number of nodes with any network degree larger than  $b$  becomes less than 1. Therefore,

$MP_{k=b} = M(1-\alpha)a^{-(1-\alpha)}b^{-\alpha} \approx 1$ , where  $M$  is the size of the network and hence:

$$b(\alpha, a) = (M(1-\alpha))^{\frac{1}{\alpha}} a^{\frac{1-\alpha}{\alpha}}.$$

**Functions of the distribution parameters Constraints**

$$k_{avg}(\alpha, a) + \frac{\sigma^2(\alpha, a)}{k_{avg}(\alpha, a)} \approx \frac{\alpha - 2}{3 - \alpha} \cdot \frac{b^{3-\alpha}(\alpha, a)}{a^{2-\alpha}}, \quad k_{\min}(\alpha, a) = a$$

The growth constraint:  $z_1 = q \frac{\alpha - 2}{3 - \alpha} \cdot \frac{b^{3-\alpha}(\alpha, a)}{a^{2-\alpha}}$ . The decline constraint:  $z_2 = qa$

**Chosen range of parameter values in the estimation procedure**  $q$  is chosen in the range between 0.0005 and 0.5.

$\alpha$  is chosen in the range between 2 and 3.

$$a = \frac{z_2}{q}$$

Note that the estimation problem turns to be two-dimensional (i.e. it is based on unrestricted choices of  $q$  and  $\alpha$ .) This is because the growth constraint yields a tangled equation for the unknown parameter  $\alpha$ .

We now present a set of three studies that evaluate the accuracy of the proposed approach.

## 4. Study I – Network Estimations Tests using Simulated Data

In this study we examine the proposed method on a large set of network degree distributions in a large number of diffusion patterns. Relying exclusively on empirical data, usually only a limited range for a defined set of variables can be tested (Casti 1996). Hence it is common to test sensitivity of such methods on simulated data, which also allows controlled tests on wider ranges of parameters than are typically available in field data. For this purpose we conducted computer simulations to generate a number of diffusion scenarios on various network structures. Each scenario produced an adoption curve that was used as input for testing the network degree distribution estimation method. We were then able to evaluate the accuracy of the estimation method by comparing the reconstructed network degree distribution with the distribution planned in the simulation. The advantage of using simulated data is that it allows testing the sensitivity of proposed methods on a wide range of cases and parameters where the true distribution is known.

Overall, we tested our method on 40 growth processes (i.e. 10 parameter sets for each of the four degree-distribution classes). We tested sets with a relatively high/low  $q$  (between 0.0005 and 0.5) and a relatively high/low  $p$  (between 0 and  $q$ ) in order to scan different dynamic scenarios (Farley et al. 1995). For each generated growth process, 10 fitting procedures were conducted (i.e. a total of 400 fits for the entire study), and the fitted adoption curve parameters with the best resulting  $R^2$  were considered the estimated solution parameter set (we also tried greater numbers of fits per generated network and found that 10 is a sufficient amount to reach best accuracy). The results of the comparisons to the known parameter sets are given in Table 2 in the form of errors in the real parameter estimations. For each parameter and network type, we took the error estimation to be the standard deviation of the distribution of errors around the real parameters. The mean and standard deviations of those errors are presented for each of the estimated parameters in Table 2.

**Table 2 Estimation Errors for the Reconstruction Method over Simulated Data**

	$\pi_1^\dagger$	$\pi_2^\dagger$	q	P
Uniform network	Mean=6.4% Std= 6.0% (a)	Mean=6.8% Std= 7.5% (b)	Mean=6.6% Std= 8.3%	Mean=8.2% Std= 9.1%
Gaussian network	Mean=5.7% Std= 7.2% ( $\mu$ )	Mean=11.8% Std= 8.5% (S)	Mean=7.9% Std= 7.3%	Mean=11.4% Std= 7.2%
Lognormal network	Mean=7.0% Std= 10.1% ( $\mu$ )	Mean=13.4% Std= 14.5% (S)	Mean=13.3% Std= 14.2%	Mean=12.4% Std= 7.1%
Scale-free network	Mean=7.3% Std= 10.4% (a)	Mean=3.4% Std= 4.5% ( $\alpha$ )	Mean=12.1% Std= 17.3%	Mean=10.4% Std= 8.2%

<sup>†</sup> All network structure parameters are defined in Table 1.

The accuracy of this method using the simulation set ranges from 6% to 14% (with a standard deviation ranging from 6% to 17%). This finding suggests that it is possible to extract the type of the network and estimate the parameters, given a single adoption pattern, to the mentioned level of accuracy, for a large range of variables. In all of the tested cases, the correct degree distribution was recovered.

However in such a simulation-based approach, the models themselves are tested using other, simple, models. Real-life phenomena include a richer set of mechanisms and noise. To address this, study 2 and 3 are presented next.

## 5. Study II – Testing Using Real-Case Data

Our evaluation of the proposed network degree estimation procedure was based exclusively on synthetic data. To address more realistic situations, where noise and other factors are involved, we tested real-life adoptions.

Since our model is mainly relevant to classical diffusion cases, we defined inclusion criteria for data sets:

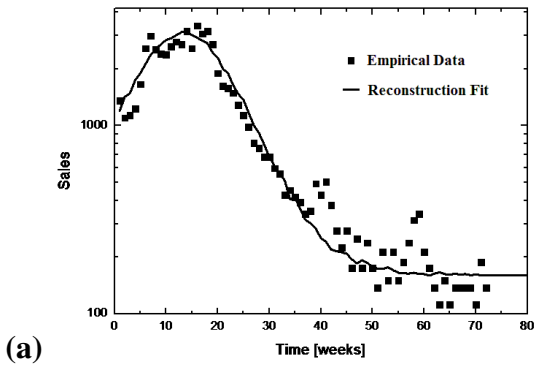
1. The growth process has an identifiable (dominant) peak, with fluctuations that are relatively small to it.
2. The time series has sufficient resolution, to allow differentiation between patterns. The pattern comprises at least 50 points of data.
3. The number of adopters is larger than several thousands of adopters, in order to allow a smooth pattern of dissemination rate.

Overall we collected 17 processes/data sets that include CD sales data, online movie penetration data (based on search query volume), petition signing rates, and adoption data of online thematic groups.

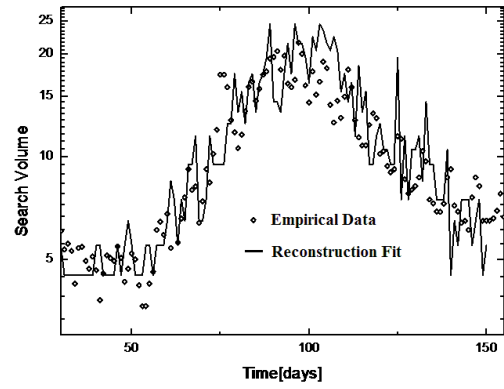
Despite the inherent noise and potentially high interference of external events in these real-life cases (e.g. external dissemination perturbations to the system, non-homogeneous campaigns, etc.), the reconstructed dissemination curves fit the data at a relatively high level of goodness-of-fit (R-Squared in the range of 90-98%). The processes in the figure are grouped by type of degree distribution of the network underlying the diffusion process.

Figure 3 includes a selection of the cases tested. Starting with light-tailed degree distributions, the uncovered networks in the first two examples exhibit normal degree distributions. In Figure 3a we present sales of an audio CD (Moe and Fader 2001) (Gaussian Pattern,  $R^2 = 0.97$ ). In Figure 3b we present the search volume (recorded by the Google Trends website, Trends 2009) for the name of the movie "Kite Runner" during its release (Gaussian pattern,  $R^2 = 0.94$ ).

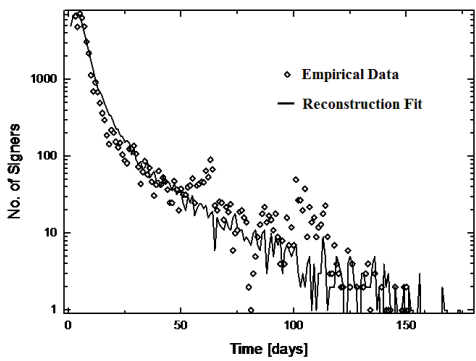
**Figure 3 Fit to Empirical Data: a) CD sales b) The movie "Kite Runner" c) Online petition site d) The movie "Cloverfield" e) Online social thematic group f) "Witty worm"**



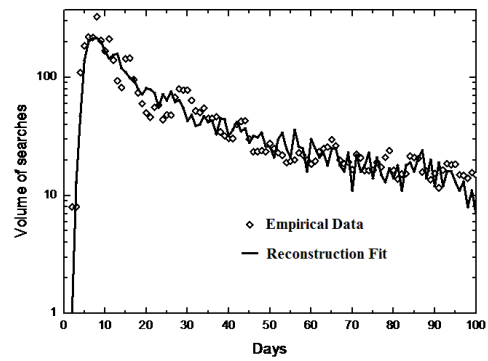
(a)



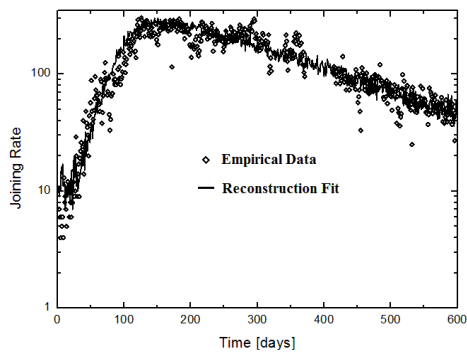
(b)



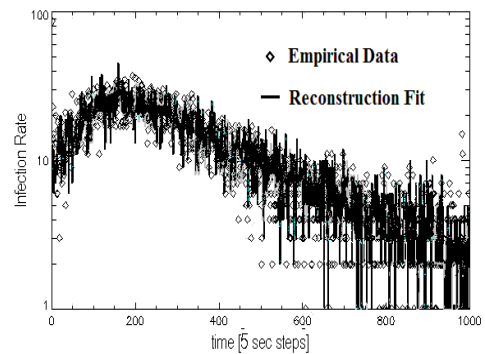
(c)



(d)



(e)



(f)

The next case, signing of online petitions, fits a uniform network degree distribution, and is given in Figure 3c (Uniform pattern,  $R^2 = 0.91$ ).

The final three examples are of heavy-tailed degree distributions: In Figure 3d we present the daily number of searches for the term "Cloverfield" (lognormal pattern  $R^2=0.98$ ), which is the name of a movie that was exceptionally credited for its efficient online viral marketing campaign (Cloverfield 2008). The pattern of these searches exhibits the pattern of a lognormal

degree distribution. Since its campaign was conducted mainly on the internet, including sites such as YouTube, the lognormal pattern is not surprising (Limpert et al 2001). The recording of a group of 74,500 users who are members of an online social network, *Friendster* (Friendster 2009) which is identified as scale free, is depicted in Figure 3e (scale-free pattern  $R^2=0.90$ ). The diffusion process records the number of users who opt to join one of the network's thematic groups, specifically, a group of fans of a certain category of TV shows. The final case is the spread of an internet worm called the "Witty Worm" (CAIDA 2004), expressed by the pattern of computer host infection rate as illustrated in Figure 3f (scale free pattern  $R^2=0.94$ ).

In addition, we took each degree distribution and conducted the estimation procedure for it as if it were the actual network. In order to test the accuracy of the identification method we introduce Table 3 where all 17 cases fits are presented. The numbers in bold are the best  $R^2$  fit results, i.e. the assumed underlying network structure, whilst the remaining three columns present  $R^2$  fit results for the remaining three network types. It is apparent for all cases that the identified network structure (i.e. degree distribution) has  $R^2$  fits that are mostly above 90% and is distinctly different from other types of structures. Interestingly, we see what appears to be a surprising number of non-scale-free active networks (more than 45% of the cases), implying that an automatic assumption that diffusion occurs over a scale-free network may be wrongly overused.

**Table 3**  $R^2$ s for all Data Sets and Networks

	<b>Scale Free</b>	Lognormal	Uniform	Gaussian
Case 1	<b>0.95</b>	0.83	0.72	0.81
Case 2	<b>0.94</b>	0.89	0.78	0.64
Case 3	<b>0.95</b>	0.72	0.88	0.68
Case 4	<b>0.91</b>	0.68	0.70	0.45
Case 5	<b>0.95</b>	0.83	0.54	0.31
Case 6	<b>0.90</b>	0.25	0.40	0.81

Case 7	<b>0.93</b>	0.75	0.64	0.22
Case 8	<b>0.94</b>	0.77	0.81	0.53
Case 9	<b>0.91</b>	0.70	0.45	0.59

---

### Identified Uniform Networks

	Scale Free	Lognormal	<b>Uniform</b>	Gaussian
Case 10	0.90	0.86	<b>0.95</b>	0.52
Case 11	0.81	0.86	<b>0.95</b>	0.57
Case 12	0.81	0.63	<b>0.91</b>	0.78

---

### Identified Gaussian Networks

	Scale Free	Lognormal	Uniform	<b>Gaussian</b>
Case 13	0.68	0.61	0.79	<b>0.94</b>
Case 14	0.66	0.78	0.86	<b>0.90</b>
Case 15	0.81	0.87	0.63	<b>0.97</b>
Case 16	0.65	0.63	0.73	<b>0.90</b>

---

### Identified Lognormal Networks

	Scale Free	<b>Lognormal</b>	Uniform	Gaussian
Case 17	0.82	<b>0.98</b>	0.71	0.4

---

While this study tests the proposed method using real data, it still is not sufficient to rule out the possibility of an alternative network structure with different process dynamics that converges to the same adoption curve. The third study was designed to address this issue.

## 6. Study III – Validation of Uniqueness versus Known Network Structure

This study aims to address two important issues. First, a more stringent test for the proposed method is to demonstrate that the estimated degree distributions are close to actual degree distributions that are known from an external source (e.g., a direct mapping of the network). A second important question is whether this procedure leads to a unique solution, or whether different degree distribution parameters or even different network categories can generate the same adoption curve with different adoption characteristics. To test the robustness of this method and the extent to which the converged network is identical to the existing network, the following empirical study is presented.

Albeit rare, data of growth processes along with the underlying networks are becoming more available these days. One such example is historical data on online social networks and their thematic groups. Users establish memberships, choosing from a wide selection of groups and categories (e.g. TV shows, local town groups, alumni groups). Group membership offers certain benefits, such as encounters with others who share the same interests or exposure to relevant information. In our study, we consider joining a group an adoption decision.

We use data from *Friendster* (Friendster, 2009), an online social network with about 100 million users. The advantage of this online social network as a data source is that the time a user joins a group is documented, and network data are largely in the public domain. We define the *active* network as the network exclusively comprising group members (i.e. all Friendster members who eventually become adopters of membership in a specific group), and compared the results generated by our proposed reconstruction method to the active network mapped on the basis of documented data.

In Table 4 we list the results of identification and estimation of the active networks for three diffusion data sets in which the active network is directly mapped (Friendster 2009). For each mapped active network (corresponding to membership in a group), the fit to four representative categories of a degree distribution (Gaussian/Poissonian, uniform, lognormal, and scale-free)

was tested. We calculated both the Cross Entropy (as used for example in Graber et al. 2004), which is a common measure of the "distance" between two distributions, and the more intuitive  $R^2$  measure, only this time between distributions.

For each network, we mark the lowest cross entropy and lowest distributions fit  $R^2$  (i.e. closest distributions) in Table 4. For the first network (denoted A), the best fit was obtained for a scale-free degree distribution. The first two moments of the reconstructed network were<sup>6</sup>  $k_{\min} = 13 \pm 2$  and  $\alpha = 2.22 \pm 0.15$ , compared to the actual values of the active network:  $k_{\text{trans}} = 12 \pm 2$  ( $k_{\text{trans}}$  is the degree at which the transition to a power-law tail takes place) and  $\alpha = 2.59 \pm 0.23$ .

The second network (B) is apparently Poissonian.<sup>7</sup> Indeed, the best fit was obtained for a Poissonian structure. The parameter of the reconstructed network was  $k_{\text{avg}} = 2$  vs.  $k_{\text{avg}} = 2$ , which is quite accurate. The third network (C) parameters are  $k_{\min} = 4 \pm 1$  and  $\alpha = 2.25 \pm 0.18$ , compared to the real values of the active network:  $k_{\text{trans}} = 5 \pm 2$  and  $\alpha = 2.66 \pm 0.4$ . The fourth network (D), which was identified as a network with a Poisson degree distribution, was shown by the proposed method to have  $k_{\text{avg}} = 5 \pm 2$  vs. the empirical  $k_{\text{avg}} = 7$ . Finally, the  $R^2$  for the distributions shows very good fits for networks A and B (0.95 and 0.98 respectively), and  $R^2$  values that are less than 90% for C and D but still high, with a much better cross entropy measure relative to the alternative fitted networks.<sup>8</sup>

#### **Table 4 Real Network Mapping Study**

---

<sup>6</sup> The estimations for the parameters of the scale-free case were calculated using numerical fits, using a method described by Newman (2005).

<sup>7</sup> This network consists of small-degree numbers, and hence the Gaussian/Poissonian class is represented by a Poissonian network degree distribution, as shown in Table 2.

<sup>8</sup> We may assume that in the case of network C and D, besides the network effect on the diffusion pattern, there were other effects at play since fits for all networks were low, relative to network A, which is also identified as scale free.

		Scale-Free	Poissonian	Lognormal	Uniform
Network A (scale-free)	Cross Entropy	<b>0.13</b>	1.20	0.92	1.13
	Distributions $R^2$	<b>0.98</b>	0.30	0.69	0.75
Network B (Poissonian)	Cross Entropy	1.44	<b>1.40</b>	1.46	1.51
	Distributions $R^2$	0.42	<b>0.95</b>	0.36	0.18
Network C (scale-free)	Cross Entropy	<b>0.22</b>	1.42	1.25	1.36
	Distributions $R^2$	<b>0.78</b>	0.18	0.51	0.45
Network D (Poissonian)	Cross Entropy	1.36	<b>1.02</b>	1.47	1.40
	Distributions $R^2$	0.54	<b>0.81</b>	0.50	0.74

Overall, the underlying network structures were correctly identified in all cases, and the errors in parameter estimation ranged between 5% and 28% (although only one estimated parameter exceeded an error of 15%).

## 7. Discussion and Managerial Applications

We have shown here how it is possible to identify the typically hidden structure of the network that actively participates in the penetration process from aggregate-level adoption data alone, and how to estimate the parameters of the degree distribution. This method of uncovering and reconstructing the active network through identification and estimation generates reasonable results (exhibiting  $R^2$ s greater than 90%), using only a single aggregate-level curve of new product penetration data on that social network.

We have also shown that different active networks can possess different types of degree distributions, even if the potential, overt network is expected to have a scale-free degree distribution.

In view of the challenge of identifying and reconstructing the active subset of the overt social network, the proposed approach may help firms more accurately assess the structure of influence affecting their potential consumers. As we have shown, disregarding the effect of network structure on the magnitude of contagion might lead to misinterpretations of the penetration data and serious biases in estimations of the penetration process. Modeling the process together with the network structure offers a more accurate estimation of the contagion process, and the efficiency of the external forces. This may also improve forecasts of future dissemination processes.

More accurate knowledge of the degree distribution may also help address managerial questions such as: how many influentials exist in a market and what are their degree centralities? What is the real effect of the network on economic market value and on individual consumers? What are the implications for information/influence flow through the market? (The degree distribution type affects the average speed of information flow, as seen in equation 9, and possibly may affect penetration levels).

One limitation of this approach is its partial applicability in the case of heavily clustered networks in which diffusion does not propagate freely over the network (i.e., in such cases, diffusion may be delayed or contained to specific clusters, Goldenberg et al. 2009b) and resulting high volatility, cyclicity and movements in the penetration pattern may obscure the dominant peak. This means that the proposed method is less relevant in cases where there is no clear dominant peak. This limitation calls for further research in this direction. For example, the method presented above may be used to investigate sub-sections of such network, and clusters that may be considered random in themselves.

## References

- Allatta, J. T., R. Iyengar, C. Van den Bulte. 2009. Social Network Dynamics after a Corporate Acquisition: How Cross-cutting Circles, Reciprocity, and Managerial Dominance Shape Networks. Working paper, University of Pennsylvania, Philadelphia, PA.
- Alon, U. 2007. Network motifs: theory and experimental approaches. *Nat. Rev. Genet.* **8**(6) 450-461.
- Amaral, L. A. N., A. Scala, M. Barthelemy, H. E. Stanley. 2000. Classes of small-world networks. *P. Natl. Acad. Sci. USA* **97**(21) 11149-11152.
- Barabasi, A. L., R. Albert. 1999. Emergence of Scaling in Random Networks. *Science* **286** (5439) 509-512.
- Barabasi, A. L., R. E. Crandall. 2003. Linked: The New Science of Networks. *Am. J. Phys.* **71**(4) 409-410.
- Barabasi, A. L. 2003. *Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life*. Plume Books, New York NY.
- Bass, F. M. 1969. A New Product Growth for Model Consumer Durables. *Manage. Sci.* **15**(5) 215-227.
- Bonabeau, E. 2002. Agent-based modeling: Methods and techniques for simulating human systems. *P. Natl Acad. Sci.* **99**(3) 7280-7287.
- Braun, M., A. Bonfrer. 2009. Censoring, Interdependence and Scalability for Dyadic Social Media. Working paper, Massachusetts Institute of Technology, Cambridge, MA.
- CAIDA. 2004. Witty Worm Data, public access (collection), CAIDA Network Telescope Project - Witty <http://imdc.datcat.org/collection/1-0019-8=CAIDA+Witty+Worm+Data+%2C+public+access..>"
- Casti, J. L. 1996. *Would-Be Worlds: How Simulation Changes the Frontiers of Science*. John Wiley & Sons, New York, NY.
- CDC. 2009. <http://www.cdc.gov/>.
- Cloverfield, Movie Marketing Madness. 2008. <http://www.moviemarketingmadness.com/blog/2008/01/17/movie-marketing-madness-cloverfield/>.
- Devore, J. L. 1995. *Probability and statistics for engineering and the sciences*. Duxbury Press, Belmont, Cal.
- Erdős, P., A. Rényi. 1959. On Random Graphs. I. *Publ. Math. Debrecen* **6** 290-297.

- Farley, J. U., D. R. Lehmann, A. Sawyer. 1995. Empirical Marketing Generalization Using Meta-Analysis. *Market. Sci.* **14** G36-G46.
- Friendster. 2009. <http://www.friendster.com>.
- Garber, T., J. Goldenberg, B. Libai, E. Muller. 2004. From Density to Destiny: Using Spatial Dimension of Sales Data for Early Prediction of New Product Success. *Market. Sci.* **23**(3) 419-428.
- Garlaschelli, D. M. I. Loffredo. 2008. Maximum likelihood: Extracting unbiased information from complex networks. *Phys. Rev. E* **78**(1) 015101.
- Goldenberg, J., S. Han, D. R. Lehmann, and J. W. Hong. 2009. The Role of Hubs in the Adoption Process. *J. Marketing* **73**(2) 1-13.
- Goldenberg, J., O. Lowengart, D. Shapira. 2009. Zooming In: Self-Emergence of Movements in New Product Growth. *Market. Sci.* **28**(2) 274-292.
- Goldenberg, J., B. Libai, E. Muller. 2001. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Market. Lett.* **12**(3) 211-223.
- Golder, P. N., G. J. Tellis. 1997. Will It Ever Fly? Modeling the Takeoff of Really New Consumer Durables. *Market. Sci.* **16**(3) 256-270.
- Gupta, S., C. F. Mela, J. M. Vidal-Sanz. 2006. The value of a "free" customer. Working paper, Harvard University, Cambridge, MA.
- Hill, S., F. Provost, C. Volinsky. 2006. Network-based Marketing: Identifying likely Adopters via Consumer Networks. *Stat. Sci.* **22**(2) 256-276.
- Iacobucci, D. 1996. *Networks in Marketing*, Sage, Thousand Oaks CA.
- Katz, E., P. F. Lazarsfeld. 1955. *Personal influence; the part played by people in the flow of mass communications*. Transaction Publishers, Glencoe, Ill.
- Katona, Z., M. Sarvary. 2008. Network Formation and the Structure of the Commercial World Wide Web. *Market. Sci.* **27**(5) 764-778.
- Katona, Z., P. Zubcsek, M. Sarvary. 2009. Network Effects and Personal Influences: Diffusion of an Online Social Network. Working paper, University of California, Berkeley, CA.
- Kossinets, G., D. J. Watts. 2006. Empirical Analysis of an Evolving Social Network. *Science* **311**(5757) 88-90.
- Krackardt, D. 1996. Structural Leverage in Marketing. Dawn Iacobucci, ed. *Networks in Marketing*, Sage, Thousand Oaks CA, 50-59.
- Liben-Nowell, D., J. Kleinberg. 2008. Tracing information flow on a global scale using Internet chain-letter data. *P. Natl. Acad. Sci. USA* **105**(12) 4633-4638.

- Limpert, E., W. A. Stahel, M. Abbt. 2001. Log-normal distributions across the sciences: Keys and clues. *Bioscience* **51**(5) 341-352.
- Mayzlin, D. 2002. The Influence of Social Networks on the Effectiveness of Promotional Strategies. Working Paper, Yale University, New Haven, CT.
- Mayzlin D., H. Yoganarasimhan. 2009. Link to Success: How Blogs Build an Audience by Monitoring Rivals. Working paper, Yale University, New Haven, CT.
- Moe, W. W., P. S. Fader. 2001. Modeling hedonic portfolio products: A joint segmentation analysis of music compact disc sales. *J. Marketing Res.* **38**(3) 376-385.
- Newman, M. E. J. 2005. Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.* **46**(5) 323-351.
- Newman, M. E. J. 2003. The structure and function of complex networks. *Siam Rev.* **45**(2) 167-256.
- Newman, M. E. J., A. L. Barabási, D. J. Watts. 2006. *The structure and dynamics of networks*. Princeton University Press, Princeton, NJ.
- Newman, M. E. J. 2002. Spread of epidemic disease on networks. *Phys. Rev. E* **66**(1) 16128.
- Nishiura, H. 2007. Time variations in the transmissibility of pandemic influenza in Prussia, Germany, from 1918-19. *Theor. Biol. Med. Model.* **4** 20-28.
- Ramasco, J. J., M. Mungan. 2008. Inversion method for content-based networks. *Phys. Rev. E* **77**(3) 12.
- Rangaswamy, A., P. Ebbes, Z. Huang. 2007. Sampling Social Networks – The Good, The Bad, and The Ugly. Presented in Marketing Science. Ann Arbor.
- Shaikh, N. I., A. Rangaswamy, A. Balakrishnan. 2006. Modeling the Diffusion of Innovations through Small-World Networks. Under Review.
- Stephen, A. T., O. Toubia. 2009. Deriving Value from Social Commerce Networks. *J. Marketing Res.* Forthcoming.
- Stumpf, M. P. H., C. Wiuf, R. M. May. 2005. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *P. Natl. Acad. Sci. USA* **102**(12) 4221-4224.
- Trends, Google. 2009. <http://www.google.com/trends/>.
- Trusov, M., W. Rand. 2009. Identifying Network Properties from Aggregate Data. Working paper, University of Maryland, Baltimore, MD.
- Valente, T. W. 1995. *Network Models of the Diffusion of Innovations*. Hampton Press, Cresskill, NJ.

Van den Bulte, C., G. L. Lilien. 1997. Bias and Systematic Change in the Parameter Estimates of Macro-level Diffusion Models. *Market. Sci.*, **16**(4) 338-353.

Van den Bulte, C., G. L. Lilien. 2001. *Medical Innovation Revisited: Social Contagion versus Marketing Effort*. *Am. J. Sociol.* **106** (March) 1409-1435.

Van den Bulte, C., Y. V. Joshi. 2007. New Product Diffusion with Influentials and Imitators. *Market. Sci.* **26**(3) 400-421.

Van den Bulte, C., S. Wuyts. 2007. *Social Networks and Marketing*. Marketing Science Institute, Cambridge, MA.

Watts, D. J., S. H. Strogatz. 1998. Collective dynamics of 'small-world' networks. *Nature* **393**(6684) 440-442.

Watts, D. J., P. S. Dodds. 2007. Influentials, Networks, and Public Opinion Formation. *J. Consum. Res.* **34**(4) 441-458.

## Appendices

### A) The dynamics of diffusion on a generalized random network

In this appendix we develop formally the diffusion dynamics on a generalized random network and then evaluate penetration evolution at early stages (growth) and late stages (decline) of the process. The probabilities of a potential adopter to be influenced by external influence (e.g. marketing efforts) and word-of-mouth communication applied by each one of her actual adopter neighbors in each time step  $\Delta t$  are  $p\Delta t$  and  $q\Delta t$ , respectively. Hence, the expected value of the number of adopters within a short time interval  $\Delta t$  after the time  $t$  is

$\Delta N(t) = \sum_x H_x(t) \cdot (p + xq)\Delta t$ . Under the continuous limit:

$$\frac{dN(t)}{dt} = \sum_x H_x(t) \cdot (p + xq), \quad (\text{A1})$$

where  $H_x(t)$  is the number of potential adopters of order  $x$  at time  $t$  defined as the number of potential adopters with exactly  $x$  actual adopters among their neighbors at time  $t$ , and the total number of potential adopters at time  $t$  is  $\sum_x H_x(t) = M - N(t)$  where  $M$  is the market potential.

At  $t = 0$  (the new product launch time), the initial conditions are  $H_0(t = 0) = M$  and

$H_x(t = 0) = 0$  for all  $x \neq 0$ . If the network is sufficiently large compared to the maximal degree in the network, the network is sparsely connected. As the network is random and hence does not contain short cycles, we can assume that within a short time interval  $\Delta t$ , the order of potential adopters cannot increase by more than 1 (i.e., the probability of simultaneous adoption of more than one potential adopter's neighbor in the same short time interval  $\Delta t$  is extremely low).

Therefore, the change in the number of potential adopters of order  $x$  at time  $t$  is given by:

$$\Delta H_x(t) = -\Delta N_x(t) - \Delta H_{xx+1}(t) + \Delta H_{x-1x}(t), \quad (\text{A2})$$

where  $\Delta N_x(t) = H_x(t) \cdot (p + xq)\Delta t$  is the number of potential adopters of order  $x$  who become actual adopters within a short time interval  $\Delta t$  after the time  $t$ , and  $\Delta H_{xx+1}(t)$  is the number of potential adopters of order  $x$  who increase their potentiality order to  $x+1$  as result of their neighbors' decision to adopt at the time  $t$ .

In general, an adoption of each potential adopter of order  $y$  and degree  $k$  increases the potentiality order of her  $k - y$  potential adopter neighbors by 1. Thus, on average, the total number of potential adopters who increase their order by 1 within a short time interval  $\Delta t$  after time  $t$  is  $\sum_y \Delta N_y(t) \sum_k (k - y) f_{kly}(t)$ , where  $f_{kly}(t)$  is the conditional probability that potential adopters of order  $y$  at time  $t$  have network degree  $k$ . By definition,  $f_{kly}(t) = 0$  for all  $y > k$  (as the number of individuals' adopter neighbors is bounded by her network degree) and the normalization condition is  $\sum_k f_{kly}(t) = 1$ . Because the network is random,  $\Delta H_{xx+1}$  (the number of potential adopters of order  $x$  that increase their order to  $x+1$ ) within a short time interval  $\Delta t$ , is proportional to the number of potential adopters of order  $x$  (excluding "perfect holes", who are potential adopters all of whose neighbors are adopters and hence can not further increase their order of potentiality). Namely,

$$\Delta H_{xx+1}(t) = \frac{H_x(t)(1 - f_{xlx}(t))}{\sum_z H_z(t)(1 - f_{zlx}(t))} \cdot \sum_y \Delta N_y(t) \sum_k (k - y) f_{kly}(t). \text{ Hence the dynamical evolution of}$$

the number of potential adopters of order  $x$  described by Equation A2 can be rewritten in the continuous limit as follows:

$$\frac{dH_x(t)}{dt} = -(p + xq + (1 - f_{xlx}(t))w(t))H_x(t) + (1 - f_{x-1lx-1}(t))w(t)H_{x-1}(t), \quad (\text{A3})$$

where

$$w(t) = \frac{\sum_y h_y(t)(p+xq) \sum_k (k-y) f_{kly}(t)}{1 - \sum_z h_z(t) f_{zly}(t)} \quad (\text{A4})$$

is the average number of potential adopters who increase their order of potentiality as a result of a single individual's adoption at the time  $t$ , and  $h_y(t) = \frac{H_y(t)}{\sum_z H_z(t)}$  is the proportion of potential adopters of order  $x$  among the entire population of potential adopters at the time  $t$ .

In order to position a closed system of dynamical equations we should also retrieve the

dynamics of the conditional probabilities  $f_{kly}(t)$ .  $f_{kly}(t) = \frac{H_x^{(k)}(t)}{H_x(t)}$ , where  $H_x^{(k)}(t)$  is the

number of potential adopters of order  $x$  with network degree  $k$ . In the case where  $H_x(t) \neq 0$ ,

$$f_{kly}(t + \Delta t) = \frac{H_x^{(k)}(t) + \Delta H_x^{(k)}(t)}{H_x(t) + \Delta H_x(t)} = \frac{f_{kly}(t) + \frac{\Delta H_x^{(k)}(t)}{H_x(t)}}{1 + \frac{\Delta H_x(t)}{H_x(t)}} \approx f_{kly}(t) + \frac{\Delta H_x^{(k)}(t) - f_{kly}(t) \Delta H_x(t)}{H_x(t)}. \quad (\text{A5})$$

For the same reasons that apply to Equation A2,

$\Delta H_x^{(k)}(t) = -\Delta N_x^{(k)}(t) - \Delta H_{xx+1}^{(k)}(t) + \Delta H_{x-1x}^{(k)}(t)$ , where  $\Delta N_x^{(k)}(t) = \Delta N_x(t) f_{kly}(t)$  is the number

potential adopters of order  $x$  and network degree  $k$  who adopt the innovation at the time  $t$ , and

$\Delta H_{xx+1}^{(k)}(t) = \Delta H_{xx+1} \tilde{f}_{kly}(t)$  is the number of potential adopters of order  $x$  and network degree  $k$

that increase their order to  $x+1$  as result of a neighbor's decision to adopt at the time  $t$ . Here,

$\tilde{f}_{kly}(t)$  denotes the conditional probability at time  $t$  that the network degree of a current

adopter's neighbor is  $k$ , given that the neighbor is potential adopter of order  $x$ . Because a

potential adopter of order  $x$  and network degree  $k$  has  $k-x$  connections with other potential

adopters (note that the current adopter was a potential adopter until the time  $t$ ), it follows,

using the distribution of a node's neighbors (Albert and Barabasi 2003), that in the case of a sufficiently large random network

$$\tilde{f}_{k|x}(t) = \frac{(k-x)f_{k|x}(t)}{\sum_{k'}(k'-x)f_{k'|x}(t)}. \quad (\text{A6})$$

Substituting the explicit expressions of  $\Delta H_x^{(k)}(t)$  and  $\Delta H_x(t)$  in Equation A5 (recall that  $\Delta N_x(t) = H_x(t) \cdot (p + xq)\Delta t$  and  $\Delta H_{x+1}(t) = (1 - f_{x|x}(t))w(t)H_x(t)\Delta t$ ) while setting to the continuous limit, we find that in the case of  $H_x(t) \neq 0$

$$\frac{df_{k|x}(t)}{dt} = \frac{H_{x-1}(t)}{H_x(t)} w(t)(1 - f_{x-1|x-1}(t))(\tilde{f}_{k|x-1}(t) - f_{k|x}(t)) - w(t)(1 - f_{x|x}(t))(\tilde{f}_{k|x}(t) - f_{k|x}(t)). \quad (\text{A7})$$

On the other hand, in the case of  $H_x(t) = 0$

$$f_{k|x}(t + \Delta t) = \frac{\Delta H_x^{(k)}(t)}{\Delta H_x(t)} = \frac{\Delta H_{x-1}^{(k)}(t)}{\Delta H_{x-1}(t)} = \tilde{f}_{k|x-1}(t), \quad (\text{A8})$$

Where, in particular, the initial conditions of the conditional probabilities are given by recursive relations. Namely, at the new product launch time the population consists of potential adopters with order of potentiality  $x = 0$  (no one has an adopter neighbor) so that  $f_{k|0}(t = 0) = P_k$  where  $P_k$  is the degree distribution of the network while for each potentiality of order  $x > 0$ ,

$$f_{k|x}(t = 0) = \lim_{\Delta t \rightarrow 0^+} f_{k|x}(\Delta t) = \tilde{f}_{k|x-1}(t = 0) = \frac{(k-x+1)f_{k|x-1}(t = 0)}{\sum_{k'}(k'-x+1)f_{k'|x-1}(t = 0)}.$$

### The growth stage.

In a random network, the probability that two neighbors of the same individual are neighbors themselves is extremely low. Thus, at relatively early stages of the process the number of potential adopters who have more than one neighbor who is an adopter of the product is very small compared to the number of potential adopters with either one or no adopter neighbor.

Furthermore, as most of the population has a network degree that is greater than 1 while the number of adopters among most individuals' neighbors is either zero or one at the initial stages of the adoption process, we can assume that  $f_{0i0}(t) \ll 1$  and  $f_{i11}(t) \ll 1$  when  $t$  is small. Thus, the penetration dynamics at early stages of the diffusion process (see Equation A1) takes the form:

$$\frac{dN(t)}{dt} = (M - N(t))p + q(H_1(t) + 2H_2(t)) + O(H_3), \quad (\text{A9})$$

where according to equation system A3:

$$\frac{dH_0(t)}{dt} = -(p + w(t))H_0(t), \quad (\text{A10})$$

$$\frac{dH_1(t)}{dt} = w(t)H_0(t) - (p + q + w(t))H_1(t), \quad (\text{A11})$$

$$\text{And } \frac{dH_2(t)}{dt} = w(t)H_1(t) + O(H_2), \quad (\text{A12})$$

where  $w(t) = h_0(t)p \sum_k k f_{k0}(t) + h_1(t)(p + q) \sum_k (k - 1) f_{k11}(t) + O(h_2)$  and

$h_x(t) = \frac{H_x(t)}{M - N(t)}$  (see Equation A4). Let  $X_1(t) = H_1(t) + 2H_2(t) = H_1(t) + O(H_2)$  so that

$$\frac{dN(t)}{dt} = (M - N(t))p + qX_1(t) + O(H_2), \quad (\text{A13})$$

where

$$\frac{dX_1(t)}{dt} = \frac{dH_1(t)}{dt} + 2\frac{dH_2(t)}{dt} = (M - X_1(t) - N(t))p \sum_k k f_{k0}(t) + X_1(t)(p + q) \left( \sum_k k f_{k11}(t) - 2 \right) + O(H_2). \quad (\text{A14})$$

or alternatively:

$$\frac{dX_1(t)}{dt} = (M - N(t))p \sum_k k f_k(t) + X_1(t) \left\{ q \left( \sum_k k f_{k11}(t) - 2 \right) - 2p \right\} + O(H_2), \quad (\text{A15})$$

where  $f_k(t) = \sum_x h_x(t) f_{k|x}(t) = \frac{M - X_1(t) - N(t)}{M - N(t)} f_{k|0}(t) + \frac{X_1(t)}{M - N(t)} f_{k|1}(t) + O(h_2)$ , the

network degree distribution among potential adopters at time  $t$ . Recall that  $f_k(t)$  denotes the ratio of the number of potential adopters with network degree  $k$  to the total number of potential adopters at time  $t$ , so that  $f_k(t) = \frac{MP_k - N^{(k)}(t)}{M - N(t)}$  where  $N^{(k)}(t)$  is the cumulative number of

actual adopters with network degree  $k$ , and  $P_k$  is the network degree distribution (thus,  $MP_k$  is the total number of individuals in the entire population with network degree  $k$ ). Because at

early stages of the process  $\frac{N(t)}{M} \ll 1$ , we can apply the approximation

$$f_k(t) = P_k \left( 1 - \frac{N(t)}{M} \right) - \frac{N^{(k)}(t)}{M} + O\left(\frac{N^{(k)}N}{M^2}\right) + O\left(\frac{P_k N^2}{M^2}\right), \text{ where } f_k(t=0) = f_{k|0}(t=0) = P_k \text{ and}$$

hence  $f_{k|1}(t=0) = \tilde{f}_{k|0}(t=0) = \frac{kP_k}{\sum_{k'} k' P_{k'}} \equiv \tilde{P}_k$ . Therefore, linearization of Equation A15 gives:

$$\frac{dX_1(t)}{dt} = Mpk_{avg} - pX_2(t) + (\tilde{Q} - 2p)X_1(t) + O\left(\frac{X_2N}{M^2}\right) + O\left(\frac{k_{avg}N^2}{M^2}\right) + O(H_2), \quad (\text{A16})$$

where  $k_{avg} = \sum_k kP_k$  and  $\tilde{Q} = q\left(\sum_k k\tilde{P}_k - 2\right)$ . The function  $X_2(t) = \sum_k kN^{(k)}(t)$  evolves

through the dynamical equation  $\frac{dX_2(t)}{dt} = \sum_k k \frac{dN^{(k)}(t)}{dt} = \sum_k k \sum_x H_x(t)(p + xq)f_{k|x}(t)$  which

can be linearized as well to give:

$$\frac{dX_2(t)}{dt} = Mpk_{avg} - pX_2(t) + (\tilde{Q} + 2q)X_1(t) + O\left(\frac{X_2N}{M^2}\right) + O\left(\frac{k_{avg}N^2}{M^2}\right) + O(k_{avg}H_2). \quad (\text{A17})$$

Thus, the temporal derivatives of equations A13, A16, and A17 generate the following system of linear and homogeneous Ordinary Differential Equations with constant coefficients:

$$\frac{d^2N(t)}{dt^2} \approx -p \frac{dN(t)}{dt} + q \frac{dX_1(t)}{dt} \quad (\text{A18})$$

$$\frac{d^2X_1(t)}{dt^2} \approx (\tilde{Q} - 2p) \frac{dX_1(t)}{dt} - p \frac{dX_2(t)}{dt} \quad (\text{A19})$$

$$\frac{d^2X_2(t)}{dt^2} \approx (\tilde{Q} + 2q) \frac{dX_1(t)}{dt} - p \frac{dX_2(t)}{dt} \quad (\text{A20})$$

and the initial conditions are  $\left. \frac{dN}{dt} \right|_{t=0} = Mp$ ;  $\left. \frac{dX_1}{dt} \right|_{t=0} = Mpk_{avg}$  and  $\left. \frac{dX_2}{dt} \right|_{t=0} = Mpk_{avg}$ . The

solution of the sub-system A19-A20 yields:

$$\frac{dX_1(t)}{dt} = A_+ e^{\lambda_+ t} + A_- e^{\lambda_- t}, \quad (\text{A21})$$

where the  $\lambda$  s are the roots of the characteristic polynomial  $\lambda^2 - (\tilde{Q} - 3p)\lambda + 2p(q + p)$ ,

such that  $\lambda_{\pm} = \frac{1}{2}(\tilde{Q} - 3p) \left( 1 \pm \sqrt{1 - \frac{8p(q+p)}{(\tilde{Q} - 3p)^2}} \right)$  and  $A_{\pm} = \pm \frac{\lambda_{\pm}}{\lambda_+ - \lambda_-} Mpk_{avg}$ . As a result,

consider the case where the mean network degree is much larger than 1 and thus  $\tilde{Q} \gg q$ , and

the aggregate level word-of-mouth effect is much stronger than the influence of the external

influence (e.g. marketing efforts) so that  $\tilde{Q} \gg p$ . One finds that  $\lambda_+ = \tilde{Q} - 3p + O\left(\frac{p+q}{\tilde{Q}}p\right)$

and  $\lambda_- = O\left(\frac{p+q}{\tilde{Q}}p\right) \ll \lambda_+$  and hence also  $A_- \ll A_+ \approx Mpk_{avg}$ . As a consequence, Equation

A21 takes the form:

$$\frac{dX_1(t)}{dt} \approx Mpk_{avg} e^{(\tilde{Q}-3p)t}. \quad (\text{A22})$$

The solution of Ordinary Differential Equation A18 following the substitution of Equation A22 (in A18) is given by:

$$\frac{dN(t)}{dt} \approx Mp \left(1 - \frac{k_{avg} q}{\tilde{Q} - 2p}\right) e^{-pt} + Mp \frac{k_{avg} q}{\tilde{Q} - 2p} e^{(\tilde{Q}-3p)t} \quad (A23)$$

(see Equation 8).

### **The decline stage.**

At the late stages of the diffusion, the majority of the population has already adopted the innovation. Hence, most of the remained potential adopters are surrounded by adopters and therefore become "perfect holes." In other words, in the final stages of the process, the order of potentiality  $x$  and the network degree  $k$  are equal for almost all potential adopters; effectively  $f_{klx}(t) \approx \delta_{kx}$ , where  $\delta_{kx}$  is Kronecker's delta. As a result, the dynamic evolutions of the number of potential adopters with network degree  $k$  and the number potential adopters of potentiality order  $x = k$  are identical and are given by reducing Equation A3 as indicated in Equation 13.

## **B) Extracting the numerical constraints from penetration data.**

We now describe the method of extracting the numerical constraints from the penetration pattern, which is the basis for the network reconstruction method described above.

### **The growth stage**

For the growth stage, the purpose is to extract the exponential slope of the adoption rate. There are several ways to approach this problem, including several discussed in Golder and Tellis (1997). We employed several methods and averaged the results of each method to minimize errors. We employed the logistic rule through fitting part of a logistic curve (only the initial Bass-like part of the curve, to prevent the results from being affected by curve asymmetry). We also used the maximum sales growth, i.e. we identified the maximum point of the sales growth (second derivative of the cumulative adoption) and regressed for the exponential slope in a log-linear space. Finally, assuming an exponential function, we used the

"returns" function  $\left(\frac{dN(t)}{N(t)}\right)$  which is actually the slope of the exponential function, in this case, the growth rate. We expect that function to be constant in the range of a constant exponential growth. We identified the straight, constant part of the function. We also estimated the value of the function (which is the exponential slope) by choosing different groups of data points out of the data set as to minimize the regression slope of the returns function (which should be zero in the constant part). The growth stage exponential slope was taken to be the average of the points' Y value for that group of points.

### **The decline stage**

To estimate the decline stage we used different groups of data points taken from the post-peak section of the data set, in which we minimized the regression for a log-linear space. We found

that the results improve dramatically when we also used the returns function  $\left(\frac{dN(t)}{N(t)}\right)$  which also shown an exponential decline towards the end of the diffusion process, coinciding with the exponential decline of the adoption rate. This is because the exponential decline of the returns function is much less noisy and lasts for a longer time.

Finally, both growth and decline slopes are used in the numerical reconstruction method described above (reflected in equations (16) and (17)).