# Sensitivity of Diffusion Dynamics to Network Uncertainty

**Abhijin Adiga, Chris Kuhlman, Henning S. Mortveit** and **Anil Kumar S. Vullikanti**
Network Dynamics and Simulation Science Laboratory,
Virginia Bioinformatics Institute,
Virginia Tech, VA 24061
{abhijin, ckuhlman, hmortvei, akumar}@vbi.vt.edu

## Abstract

Simple diffusion processes on networks have been used to model, analyze and predict diverse phenomena such as spread of diseases, information and memes. More often than not, the underlying network data is noisy and sampled. This prompts the following natural question: how sensitive are the diffusion dynamics and subsequent conclusions to uncertainty in the network structure?

In this paper, we consider two popular diffusion models: *Independent cascades* (IC) model and *Linear threshold* (LT) model. We study how the expected number of vertices that are influenced/infected, given some initial conditions, are affected by network perturbation. By rigorous analysis under the assumption of a reasonable perturbation model we establish the following main results. (1) For the IC model, we characterize the susceptibility to network perturbation in terms of the critical probability for phase transition of the network. We find the expected number of infections is quite stable, unless the the transmission probability is close to the critical probability. (2) We show that the standard LT model with uniform edge weights is relatively stable under network perturbations. (3) Empirically, the transient behavior, i.e., the time series of the number of infections, in both models appears to be more sensitive to network perturbations. We also study these questions using extensive simulations on diverse real world networks, and find that our theoretical predictions for both models match the empirical observations quite closely.

## Introduction

A number of diverse phenomena are modeled by simple diffusion processes on graphs, such as the spread of epidemics (Newman 2003), viral marketing (Kempe, Kleinberg, and Tardos 2005; Goldenberg, Libai, and Muller 2001) and memes in online social media (Romero, Meeder, and Kleinberg 2011; Bakshy et al. 2011). It is common to associate with each vertex a state of 0 (denoting "not infected" or "not influenced") or state 1 (denoting "infected" or "influenced") in these models; each neighbor of a node in state 1 switches to state 1 based on a probabilistic rule. We focus on two such models, referred to as the *independent cascades* (IC) model (which is a special case of the SIR process), and *linear threshold* (LT) model. In most applications, however, the underlying networks are inherently noisy and incomplete, since they are often inferred by indirect measurements, for instance: (i) networks based on Twitter data (e.g., (Gonzlez-Bailn et al. 2011; Bakshy et al. 2011; Galuba 2010)) are constructed by limited samples available through public APIs, (ii) biological networks are inferred by experimental correlations, e.g., (Hagmann 2008; Schwab et al. 2010), which might be incomplete, and (iii) the Internet router/AS level graphs are constructed using traceroutes, e.g., (Faloutsos, Faloutsos, and Faloutsos 1999), which are known to give a biased and incomplete structure (see, e.g., (Achlioptas et al. 2009)).

This raises a fundamental issue for diffusion processes on networks: *How does the uncertainty in the network affect the conclusions drawn from a study of the diffusion dynamics?* For instance, how robust is an inference that there will be no large outbreak in the network, in the face of noise/uncertainty in the network? Recent statistical and simulation based studies involving perturbation of the network by "rewiring" pairs of edges (which preserves the degree sequence) show that changes in the network structure significantly alter the dynamics, and the efficacy of intervention mechanisms, even when aggregate structural properties, such as the degree distribution and assortativity are preserved (Eubank 2010; Chen 2010). Surprisingly, there is limited mathematically rigorous work to explain the empirical findings in a systematic manner, despite a large body of research on diffusion models.

Our work is motivated by these considerations of sensitivity of the dynamics to noise and the adequacy of sampling of a network $G = (V, E)$. Since there is very limited understanding of how noise should be modeled, we consider a simple *Random Edge Perturbation* model for noise, in which each pair $u, v$ of vertices is selected for addition/deletion with probability $\frac{\epsilon}{n}$, where $\epsilon > 0$ is a parameter, and $n$ is the number of vertices; thus, on average, only $\epsilon n$ edges are altered. This model has been used quite extensively both in social network analysis and computer science for understanding the sensitivity to graph properties, e.g., (Costenbader and Valente 2003; Borgatti, Carley, and Krackhardt 2006; Flaxman and Frieze 2004; Flaxman 2007). Let $R(\epsilon)$ denote the random set of edges selected by this process; we denote the perturbed graph by $G \oplus R(\epsilon)$. We study how the expected number of infections, given some initial conditions, is affected by the extent of perturbation, $\epsilon$.

**Our contributions.**

**1. The Independent cascades model**. We consider networks $G$ which exhibit a phase transition in their component sizes, with a critical probability $p_c$. In Theorem 1, we characterize the expected number of infections in the perturbed graph in terms of $p_c$: when $p < p_c$, we show that there exists a threshold $\epsilon_t = \Theta(1/p_c)$ such that for $\epsilon < \epsilon_t$, the expected number of infections in the perturbed graph remains close to that in $G$; however, for $\epsilon > c'\epsilon_t$ for a constant $c'$, there is a phase transition, and the expected number of infections after perturbation is much larger than that in $G$. The main implication is that the dynamics are quite robust to perturbations, unless the transmission probability is close to $p_c$. We find this to be consistent with extensive simulations on a large number of real networks—the sensitivity to perturbations is maximized at a point which approximately matches the threshold $\epsilon_t$ in many networks. We also examine the transient behavior (i.e., the time series of the number of infections), and find it to be more sensitive than the expected total number of infections.

**2. The Linear threshold model**. In Theorem 2, we show formally that for any network $G$ with maximum degree $D = O(n/\log n)$, the expected number of infections after perturbation, starting at $s$ random initial infections, is bounded by $O(s(D + \epsilon + \log n)\log n)$. This implies that the dynamics is quite stable for low $s$ and $\epsilon$. Our result is based on the analysis of the random graph model in which each node selects a random in-edge (which is shown to "correspond" to the LT model by (Kempe, Kleinberg, and Tardos 2003)). We first show that the diameter is bounded by $O(D' \log n)$, where $D'$ is the maximum degree of the perturbed graph, and then prove that the expected number of infections, starting at a random source, is bounded by the diameter. Our theoretical bounds corroborate well with our experimental observations on a large class of real networks, which show a gradual variation with $\epsilon$. We find that the expected number of infections grows more sharply with $\epsilon$, as the number of sources is increased. Further, as in the IC model, we find the transient behavior is more sensitive to $\epsilon$.

**Discussion and implications**. From the point of view of dynamical system theory, our work may be regarded as a study of *stability* of dynamics over a network with respect to the edge structure. The existence of the critical value for the parameter $\epsilon$ in the IC model can be thought of as a bifurcation point. Admittedly, our results only hold for the specific random edge perturbation model of noise; uncertainty in networks is a much more complex process, and might involve dependencies arising out of the network evolution. Although we focus on specific dynamical properties and the random edge perturbation model, our results give the first rigorous theoretical and empirical analysis of the noise susceptibility of these diffusion models. Further, our analytical and empirical techniques, based on the random graph characterization, are likely to help in the analysis of other, more complex, noise models, which take dependencies into account.

**Organization**. Because of space limitations, we omit the details of some of our proofs and experimental results; these will be available in the complete version of the paper.

## Related work

Noise and issues of sampling are well recognized as fundamental challenges in complex networks, and there has been some work on characterizing it and the sensitivity to different parameters, especially in network properties, such as: (i) (Costenbader and Valente 2003; Borgatti, Carley, and Krackhardt 2006) show that certain centrality measures are robust to random edge and node perturbations, and (ii) (Achlioptas et al. 2009) show that there is an inherent bias in traceroute based inference of the Internet router network, which might give incorrect degree distributions. Flaxman and Frieze (Flaxman and Frieze 2004; Flaxman 2007) formally characterize conditions under which the graph expansion and diameter is highly sensitive to random edge additions; these are among the few analytical results of this type. Some of the approaches to address noise include: (i) the prediction of missing links using clustering properties, e.g., (Clauset, Moore, and Newman 2008), and (ii) approaches such as "property testing" algorithms, e.g., (Ron 2010) and "smoothed analysis", e.g., (Spielman 2009) for efficient computation of graph properties.

To our knowledge, most work on the sensitivity of graph dynamical systems to noise in the network is empirical. However, for regular networks such as rings, topics such as synchronization and bifurcations have been studied (Kaneko 1985; Wu 2005). As discussed earlier, (Eubank 2010; Chen 2010) study the effect of changes in the network by edge rewirings on the epidemic properties. (Lahiri et al. 2008) study the effect of stochastic changes in the network on influence maximization problems. They find, using simulations, that in the LT model, the spread size is quite robust; our techniques help explain some of these observations.

## Preliminaries

**Noise models** There is no consensus on the best way to model uncertainty/noise, and we consider a simple model of random edge additions that has been studied quite extensively in social network analysis (Costenbader and Valente 2003; Borgatti, Carley, and Krackhardt 2006); some problems have also been studied analytically in this model (Flaxman and Frieze 2004; Flaxman 2007). Let $G = (V, E)$ be the unperturbed graph. Let $R(\epsilon) = (V, E(\epsilon))$ be a random graph on $V$ in which each pair $u, v \in V$ is connected with probability $\frac{\epsilon}{n}$. The perturbation graph $G' = G \oplus R(\epsilon)$ is a graph constructed in the following manner: each pair $u, v \in V$ is connected in $G'$ if $(u, v) \in R(\epsilon) - E$ or $(u, v) \in E - R(\epsilon)$. In other words, each pair $u, v$ is selected for addition/deletion with probability $\frac{\epsilon}{n}$. We also consider perturbations involving just addition of edges: this is denoted by $G + R(\epsilon)$, and consists of all edges $(u, v) \in E \cup R(\epsilon)$.

**Network diffusion models**. Let $G = (V, E)$ denote an undirected network. Here we define the diffusion models we study. In each model, each vertex $v \in V$ can be in state $x_v \in \{0, 1\}$, with state 0 denoting "inactive/uninfected/uninfluenced" and state 1 denoting "active/infected/influenced", depending on the application. We restrict ourselves to *monotone* or progressive processes, i.e., an infected node stays infected. Each node is associated with an activation function whose inputs

include the states of its neighbors. This function computes the next state of the node. The diffusion process starts with a few vertices becoming active/infected; we refer to this set as the *initial set* or the seed set. For an initial set of active nodes $S$, let $\sigma(S)$ denote the expected number of active nodes at termination; these models always reach fixed points. We consider the following models:

*1. Independent Cascade (*IC*) Model* (Kempe, Kleinberg, and Tardos 2003). This model is a special case of the SIR model for epidemics. An infected node $v$ infects each neighbor $w$ with probability $p$. Equivalently, each edge $(v, w)$ can be *live* with probability $p$, independently of all other edges. All those nodes which are connected to the initial set through a *live path* are considered infected. In the perturbed graph $G' = G + R(\epsilon)$, suppose $(v, x)$ is a newly added edge, then, $v$ tries to infect $x$ with probability $p$ and vice versa.

*2. Linear Threshold (*LT*) Model.* (Kempe, Kleinberg, and Tardos 2005) Each node $v$ is associated with a threshold $\theta_v \in [0, 1]$, chosen uniformly at random. $v$ is influenced by its neighbor $w$ according to weight $b_{v,w}$ such that $\sum_{w \in N(v)} b_{v,w} \leq 1$. Node $v$ becomes infected if $\sum_{w \in A(v)} b_{v,w} \geq \theta_v$, where $A(v) \subseteq N(v)$ is the set of neighbors of $v$ which are currently infected. In our analysis and experiments, we assume that $b_{v,w} = 1/deg(v, G)$ for all $w \in N(v)$, where $deg(v, G)$ is the degree of $v$ in $G$. This means, $v$ is influenced equally by all its neighbors. This model was considered in (Kempe, Kleinberg, and Tardos 2003). In the perturbed graph $G' = G + R(\epsilon)$, $b_{v,w} = 1/deg(v, G')$, where $deg(v, G')$ is the new degree of $v$.

## Analyzing the sensitivity of the independent cascade model

We now discuss the sensitivity of the IC model for graphs that exhibit a phase transition, which is discussed informally here. Given a graph $G$ with $n$ vertices, let $G(p)$ denote the random spanning subgraph of $G$ obtained by retaining each edge of $G$ independently with probability $p$. Many graphs (e.g., the complete graph, random regular graphs) exhibit the following property: there is a critical probability $p_c$ such that if $p < p_c$, all components in $G(p)$ are "small", namely of size $o(n)$, while if $p > p_c$, there exists a giant component of size $\theta(n)$. Similar threshold phenomena has been observed (empirically) in the real-world graphs which we study. Let $N$ denote the number of components in $G(p)$ when $p < p_c$. We note that if the number of nodes in $G$ of degree $\leq d$ is at least $cn$ for a constant $d$ (a property satisfied by scale free networks), then $N = \theta(n)$. This follows from the fact that the expected number of isolated vertices in $G(p)$ is $\geq c(1-p)^d n = c'n$, under this assumption. Using Chebychev inequality, it can be shown that with high probability, the number of isolated vertices is very close to the expected value. Hence, $N \geq c'n$ with high probability.

**Theorem 1.** *Consider the IC model on a family of graphs $G$ exhibiting the following properties: (1) it undergoes a phase transition with critical probability $p_c$, with the additional assumption that for $p < p_c$, all components in $G(p)$ are $o(\sqrt{n})$, with high probability and (2) There is a function $N = N(n, p)$, such that the number of components on percolation in $G$ at probability $p$ is within $(N, (1+\mu)N)$, with probability $1 - o(1)$, for a constant $\mu > 0$. Let $G' = G + R(\epsilon)$ denote the perturbed graph. If $p < p_c$, then, there is a threshold perturbation factor $\epsilon_t = \frac{N}{pn}$, such that for (i) $\epsilon < \epsilon_t$, the expected number of infections in $G'$ starting at a random initial node is $o(n)$, and for (ii) $1/p > \epsilon > 2(1+\delta)\epsilon_t$, for any constant $\delta > 0$, the expected number of infections in $G'$ starting at a random initial node is $\Theta(n)$ as $n \to \infty$.*

*Proof.* Let $\{C_i | i \in N\}$ be the set of connected components of $G(p)$. Let $n_i$ denote the size of $C_i$. The probability that components $C_i$ and $C_j$ are connected by at least one edge is at most $\frac{n_i n_j \epsilon p}{n}$ in $G'(p)$. Consider an instance $H$ of the Chung-Lu random graph model (Chung and Lu 2002) with $N$ nodes with weights $w_1, \ldots, w_N$, such that $w_i = n_i \epsilon p$. The probability of edge $(i, j)$ in $H$ equals $\frac{w_i w_j}{\sum_k w_k} = \frac{(n_i \epsilon p) \cdot (n_j \epsilon p)}{\sum_k n_k \epsilon p} = \frac{n_i n_j \epsilon p}{n} \geq \Pr[C_i \text{ and } C_j \text{ are connected in } G'(p)]$, since $\sum_{k=1}^{N} n_k = n$. Thus, the connectivity in the Chung-Lu instance $H$ dominates that in $G'$. The average degree $w_{avg}$ for $H$ is $w_{avg} = \sum_i \frac{w_i}{N} = \sum_i \frac{n_i \epsilon p}{N} = \frac{n \epsilon p}{N}$. From the connectivity threshold in the Chung-Lu model, it follows that $H$ has no giant component if $w_{avg} < 1$, which gives $\epsilon < \frac{N}{pn} = \epsilon_t$.

Next, suppose $1/p > \epsilon > \epsilon_t$. By inclusion-exclusion, it follows that the probability that components $C_i$ and $C_j$ are connected in $G'(p)$ by at least one edge is at least $\frac{n_i n_j \epsilon p}{n} - \frac{(n_i n_j \epsilon p)^2}{2n^2} \geq \frac{n_i n_j \epsilon p}{2n}$, because of the assumption that $n_i = o(\sqrt{n})$ and $\epsilon p < 1$. Next, consider another instance $H$ of the Chung-Lu model with $N$ nodes with weights $w_1, \ldots, w_N$, such that $w_i = n_i \epsilon p/2$. The probability of edge $(i, j)$ in $H$ equals $\frac{w_i w_j}{\sum_k w_k} = \frac{(n_i \epsilon p/2) \cdot (n_j \epsilon p/2)}{\sum_k n_k \epsilon p/2} = \frac{n_i n_j \epsilon p}{2n} \leq \Pr[C_i \text{ and } C_j \text{ are connected in } G'(p)]$. Thus, the connectivity in the Chung-Lu instance $H$ is dominated by that in $G'$. The average degree $w_{avg}$ for $H$ is $w_{avg} = \sum_i \frac{w_i}{N} = \sum_i \frac{n_i \epsilon p}{2N} = \frac{n \epsilon p}{2N}$. From the connectivity threshold in the Chung-Lu model, it follows that $H$ has a giant component if $w_{avg} > 1 + \delta$, for any constant $\delta > 0$, which gives $\epsilon > \frac{2(1+\delta)N}{pn} = 2(1+\delta)\epsilon_t$. Therefore, with high probability $G'$ has a component with $\Theta(n)$ vertices. Since there is a constant probability that the seed belongs to the giant component, it follows that the expected number of infections in this case is $\Theta(n)$. □

**Remark 1.** *We note that if $\epsilon < \epsilon_t$, for any seed set of size $s$ (not necessarily random), the expected number of infections in $G'$ is $o(sn)$.*

## Analyzing the sensitivity of the linear threshold model

We now analyze the impact of edge perturbations on the LT model on a graph $G = (V, E)$. Recall that in the specific version of the LT model we consider here, we set $b_{v,w} = 1/deg(v)$ for each node $v \in V$ and neighbor $w \in N(v)$. (Kempe, Kleinberg, and Tardos 2003) show that the fixed points and the number of infected nodes

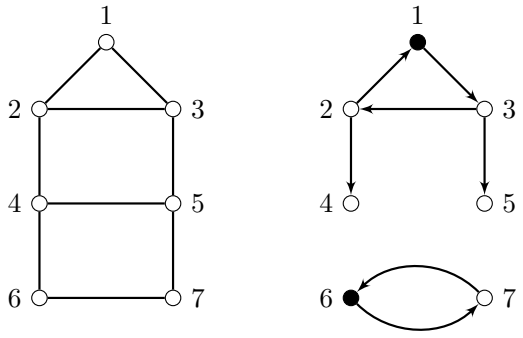



Figure 1: A graph and an instance of the random graph $H_{LT}$ corresponding to the LT model. For the component $T$ induced by $\{1,2,3,4,5\}$, 1 is chosen as the root and as a result, $T_0 = \{1\}, T_1 = \{3\}, T_2 = \{2,5\}$ and $T_3 = \{4\}$.

they have, can be studied through an elegant random graph model, which we describe here. Construct a random directed graph $H_{LT} = (V, E')$ in the following manner: For each node $v \in V$, a neighbor $w$ is chosen with probability $b_{v,w}$ and a directed edge is added from $w$ to $v$. Figure 1 illustrates a graph $G$ and an instance of $H_{LT}$. Note that even though $G$ is undirected, $H_{LT}$ is a directed graph. For a set $S \subset V$, let $\sigma(S, H_{LT})$ denote the number of nodes reachable from $S$ in $H_{LT}$ (including those in $S$). Then, (Kempe, Kleinberg, and Tardos 2003) show that $\sigma(S)$, the expected number of infections with a starting set $S$, satisfies $\sigma(S) = \sum_{H_{LT}} \Pr[H_{LT}]\sigma(S, H_{LT})$. We use this characterization to analyze the impact of edge perturbations.

The random graph $H_{LT}$ constructed by the above process has the following structure: In each connected component $T$ of $H_{LT}$, every vertex has one incoming edge and therefore, there exists exactly one directed cycle. If we choose a vertex in the cycle as the *root* $r$ and remove its incoming edge, then, $T$ remains connected and corresponds to a tree rooted at $r$ with all edges oriented away from $r$. In the rest of this section, we loosely refer to such a component as a "tree" with one cycle or sometimes just tree. $T$ can be partitioned into sets $T_0, \ldots, T_k$ such that for each $i > 0$, a vertex $v \in T_i$ has an incoming edge from some vertex $u \in T_{i-1}$. The set $T_0$ is a singleton consisting of the root vertex $r$. The incoming edge for $r$ is from some neighbor in $\cup_{i=1}^{k} T_i$. All of this is illustrated in Figure 1. First, we show the following:

**Lemma 1.** *In the LT model, let $\delta = \min_{v \in V, w \in N(v)} b_{v,w}$. Each tree in the random subgraph $H_{LT}$ has depth $O\left(\frac{1}{\delta}\log n\right)$, with probability at least $1 - \frac{1}{n^3}$.*

*Proof.* Consider a tree $T$ in $H_{LT}$, which is partitioned into sets $T_0, \ldots, T_k$, as mentioned above. For any $i = 1, \ldots, k-1$, a vertex $v \in T_i$ would become a root if it chooses an incoming edge from one of its descendants. The probability of this event is at least $\min_{w \in N(v)} b_{v,w} \geq \delta$. Therefore, the probability that none of the vertices in $T_i$ becomes a root is at most $1 - \delta$, which in turn implies that the probability that none of the vertices in $T_i, i = k-1, \ldots, 1$ becomes a root is at most $(1-\delta)^{k-2}$. Hence, the probability that $T$ has depth more than $k = c \cdot \frac{1}{\delta}\log n + 2$ for a constant $c$ is at most $\sum_{k \geq c \cdot \frac{1}{\delta}\log n + 2}^{n} (1-\delta)^{k-2} \leq \frac{1}{n^4}$. Since there are at most $n$ such trees in $H_{LT}$, the probability that any of these has depth more than $O\left(\frac{1}{\delta}\log n\right)$ is at most $\frac{1}{n^3}$. □

Consider a vertex $v$ contained in a tree $T$. Let $n(v,T)$ denote the number of vertices reachable from $v$ in $T$. Then, the number of infections resulting from $v$ is the expected value of $n(v,T)$, averaged over all random subgraphs $H_{LT}$ and trees containing $v$. Define $A(T)$ as $A(T) = \frac{1}{|T|}\sum_{v \in T} n(v,T)$. Conditioned on a random subgraph $H_{LT}$, the average number of infections starting at a random source is $\sum_{T \in H_{LT}} A(T)\frac{|T|}{n}$; the average number of infections starting at a random source is $\sum_{H_{LT}} \Pr[H_{LT}] \sum_{T \in H_{LT}} A(T)\frac{|T|}{n}$.

**Lemma 2.** *For each tree $T$ in a random subgraph $H_{LT}$, $A(T) \leq 2d$, where $d$ is the depth of $T$.*

*Proof.* Define $\hat{T}$ to be the tree obtained by removing the incoming edge for the root in $T$. As described above, $\hat{T}$ is an out-tree. For each $v \in \hat{T}$, we define $n(v,\hat{T})$ as the number of vertices reachable from $v$ in $\hat{T}$— this corresponds to the size of the subtree rooted at $v$ in $\hat{T}$. We define $A(\hat{T}) = \frac{1}{|\hat{T}|}\sum_{v \in \hat{T}} n(v,\hat{T})$, and prove that $A(\hat{T}) \leq d$. We prove this by induction on the depth of the out-tree. The base case is a leaf node $u$, for which $A(u) = 1$.

Let $r$ be the root of $\hat{T}$. Suppose it has children $v_1, \ldots, v_a$. Let $\hat{T}_i$ be the subtree rooted at $v_i$, and let $n_i$ be the number of vertices in $\hat{T}_i$. By induction, $A(\hat{T}_i) = \frac{1}{n_i}\sum_{v \in \hat{T}_i} n(v,\hat{T}_i) \leq d-1$.

$$\begin{aligned} A(\hat{T}) &= \frac{1}{|\hat{T}|}\sum_{v \in \hat{T}} n(v,\hat{T}) \\ &= \frac{1}{|\hat{T}|}n(r,\hat{T}) + \sum_{i=1}^{a}\frac{1}{|\hat{T}|}\sum_{v \in \hat{T}_i} n(v,\hat{T}_i) \\ &= 1 + \sum_{i=1}^{a}\frac{n_i}{|\hat{T}|}A(\hat{T}_i) \leq 1 + \sum_{i=1}^{a}\frac{n_i}{|\hat{T}|}(d-1) \\ &\leq 1 + \frac{|\hat{T}|-1}{|\hat{T}|}(d-1) \leq d \end{aligned}$$

The third equality follows because $n(r,\hat{T}) = |\hat{T}|$, and by definition, $A(\hat{T}_i) = \frac{1}{n_i}\sum_{v \in \hat{T}_i} n(v,\hat{T}_i)$. The first inequality follows by the induction hypothesis, since the depth of each $\hat{T}_i \leq d-1$. The second inequality follows because $\sum_{i=1}^{a} n_i = |\hat{T}| - 1$.

Next, we consider $A(T)$. We recall that $T$ is a tree with a cycle of length at most $d$. Let the cycle consist of vertices $u_0 = r, u_1, \ldots, u_b$, with $b \leq d-1$. For each $u_i$, $n(u_i,T) = |T|$, since there is a path from $u_i$ to $r$. For every other vertex $u \neq u_i$ in $T$, $n(u,T) = n(u,\hat{T})$. This implies, $A(T) \leq \frac{d|T|}{|T|} + A(\hat{T}) \leq 2d$. □

Finally, we bound the number of infections in the perturbed graph below; empirically, we find that the expected number of infections in the LT model is not very sensitive to $\epsilon$, which is consistent with the bound below, which is linear in $\epsilon$.

**Theorem 2.** *Let $G(V, E)$ be a graph with maximum degree $D$. For the LT model where $b_{v,w} = 1/deg(v)$ for each node $v \in V$ and $w \in N(v)$, the expected number of infected vertices starting with a initial random seed set of size $s$ in the perturbed graph $G + R(\epsilon)$ is $O(s(D + \epsilon + \log n) \log n)$.*

*Proof.* By a direct application of Chernoff bound, it can be shown that with probability at least $1 - \frac{1}{n^3}$, the maximum degree in $G' = G + R(\epsilon)$ is at most $D + \epsilon + c \cdot \log n$ for a constant $c$; with the remaining probability of $\frac{1}{n^3}$, the maximum degree is $O(n)$. We consider the random graph process to generate a subgraph $H_{LT}$ of $G'$. Since $b_{v,w} = 1/deg(v)$ for each node $v \in V$ and $w \in N(v)$, for this model, the value of $\delta$ of Lemma 1 is $1/D$ and therefore, each tree in $H_{LT}$ has depth at most $O((D + \epsilon + \log n) \log n)$, with probability at least $1 - \frac{1}{n^3}$. Conditioned on $H_{LT}$ satisfying this bound on the depth, $A(T) = O((D + \epsilon + \log n) \log n)$ for all $T \in H_{LT}$. For $H_{LT}$ that does not satisfy the depth bound, we have $A(T) = O(n)$ for all $T \in H_{LT}$. Therefore, the expected number of infections for a single random seed is $O((D+\epsilon+\log n) \log n)+O(\frac{n}{n^3}) = O((D+\epsilon+\log n) \log n)$. Hence proved. □

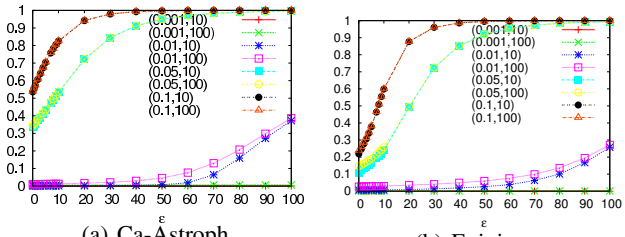
## Experimental results

We study the sensitivity to edge perturbations empirically on twenty diverse real-world networks (from (Leskovec 2011)) with varying degrees of perturbation and other factors for both IC and LT models. Our main conclusions are the following:

1. *Sensitivity in the IC model*: we find that our empirical results match quite well with Theorem 1— the expected number of infections IC model is well-behaved with $\epsilon$, unless $p$ is close to $p_c$. Further, in most networks, the sensitivity is maximized at a point which approximately matches the threshold $\epsilon_t$. Though Theorem 1 strictly holds for graphs showing a phase transition, we find that most of the networks we study exhibit such a phenomenon.
2. *LT model*: we find that the expected number of infections in the LT model is not very sensitive to $\epsilon$, especially for low number of seeds (e.g., less than 10), confirming the general bound derived from Theorem 2. When the number of seeds is large, the rate of increase of the expected number of infections seems to be higher initially.
3. *Sensitivity of transients/temporal characteristics*: our preliminary results suggest that the transient behavior (the time series of #infections versus time) is more sensitive than the expected #infections to $\epsilon$, in both models.
4. *Additions vs deletions*: we find that perturbations involving both edge additions and deletions do not alter the results by much, compared to perturbations involving just edge additions. This follows from the sparsity of the graphs, and corroborates our analytical results, to some extent.

Because of space limitations, we only discuss a sample of the results, and omit the results involving edge deletions; the remaining will be available in the complete version of the paper. We consider twenty different networks from (Leskovec 2011), with values of $\epsilon$ ranging from 0 to 100, with results averaged over 10 independent simulation runs. A simulation runs consists of 100 separate diffusion instances on one graph instance. A diffusion instance is a computation of the state of every node as a function of time, from time t=0 to the specified maximum time.

**The Independent Cascades Model**. Figure 2 shows the the expected fraction of infected nodes vs. $\epsilon$ for two networks (namely, the astrophysics co-authorship and epinions networks)— they both show low sensitivity for a broad range of $\epsilon$ values. For each of the settings, the expected number of infections rises sharply; further, the networks show differences in the plots for different parameter values. Some of the results for other networks are summarized in Table 1, which shows two sets of results for each network. Both are estimates of $\epsilon_t$, the threshold perturbation factor, using two methods. (i) In Method 1 we estimate $N$, the number of connected components in a random subgraph from the simulations, and use Theorem 1 to determine $\epsilon_t = N/np$. (ii) In Method 2, we consider the plot of infection size with respect to $\epsilon$ for a particular transmission probability $p$ (as in Figure 2), and choose $\epsilon_t$ to be the point of maximum slope of the curve on the $X$-axis. We note that both methods seem to give similar estimates of $\epsilon_t$. We empirically observe that the standard plot of infection size vs. transmission probability $p$ for all the networks (without perturbations) exhibits some kind of phase transition; these results are omitted here because of space.




(a) Ca-Astroph (b) Epinion

Figure 2: Expected #infections vs $\epsilon$ in the IC model for different pairs of transmission probability $p$ and seed set size, with the seed nodes being chosen randomly.

**Linear Threshold Model.** Figure 3 shows the expected number of infections for two representative networks— the slashdot and wiki networks. They both seem to follow the general bounds of Theorem 2. We have also studied the LT model on all the 20 networks, as in the IC model; these are omitted here because of space limitations. Figure 4 shows the sensitivity in the transient behavior, i.e., the fraction of infections by time for the LT model— as mentioned earlier, this shows a greater sensitivity to $\epsilon$.

| Network | $n$ | $\|E\|$ | $\epsilon_t$ by Method 1 | | | | $\epsilon_t$ by Method 2 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | p=0.001 | 0.01 | 0.05 | 0.1 | p=0.001 | 0.01 | 0.05 | 0.1 |
| Synthetic graphs | | | | | | | | | | |
| Regular ($d = 20$) | 10000 | 100000 | 990.0 | 90.1 | 10.0 | 1.5 | $> 100$ | 90 | 10 | 8 |
| Autonomous Systems | | | | | | | | | | |
| As20000102 | 6474 | 12572 | 998.0 | 98.1 | 18.1 | 8.2 | $> 100$ | 90 | 20 | 8 |
| Oregon1010331 | 10670 | 22002 | 997.9 | 97.8 | 17.9 | 8.078 | $> 100$ | 90 | 20 | 8 |
| Oregon2010331 | 10900 | 31180 | 997.1 | 97.1 | 17.3 | 7.715 | $> 100$ | 90 | 20 | 8 |
| Co-authorship | | | | | | | | | | |
| Astroph | 17903 | 196972 | 989.0 | 89.0 | 11.6 | 4.1 | $> 100$ | 80 | 3 | 0 |
| Condmat | 21363 | 91286 | 996.0 | 95.7 | 15.8 | 6.1 | $> 100$ | 90 | 10 | 4 |
| Grqc | 4158 | 13422 | 997.0 | 96.8 | 16.9 | 7.2 | $> 100$ | 90 | 10 | 3 |
| Hepph | 11204 | 117619 | 989.0 | 90.3 | 13.8 | 5.5 | $> 100$ | 90 | 8 | 0 |
| Hepth | 8638 | 24806 | 997.0 | 97.2 | 17.1 | 7.2 | $> 100$ | 90 | 10 | 5 |
| Citation | | | | | | | | | | |
| HepPh | 34546 | 420877 | 988.0 | 87.7 | 10.1 | 3.1 | $> 100$ | 70 | 4 | 0 |
| HepTh | 27770 | 352285 | 988.0 | 87.5 | 10.6 | 3.5 | $> 100$ | 80 | 0 | 0 |
| Communication | | | | | | | | | | |
| Email-Enron | 33696 | 180811 | 995.0 | 95.1 | 16.1 | 6.8 | $> 100$ | 90 | 10 | 0 |
| Email-EuAll | 265214 | 364481 | 999.0 | 98.6 | 18.7 | 8.8 | $> 100$ | 90 | 20 | 9 |
| Social | | | | | | | | | | |
| Epinion | 75877 | 405739 | 995.0 | 94.8 | 16.5 | 7.3 | $> 100$ | 90 | 10 | 7 |
| Slashdot0811 | 77360 | 469180 | 994.0 | 93.9 | 15.6 | 6.6 | $> 100$ | 90 | 10 | 6 |
| Slashdot0902 | 82168 | 504230 | 993.8 | 93.8 | 15.5 | 6.5 | $> 100$ | 90 | 9 | 6 |
| Twitter | 22405 | 59898 | 997.0 | 97.3 | 17.5 | 7.8 | $> 100$ | 90 | 10 | 8 |
| Wiki-Vote | 7066 | 100736 | 985.0 | 86.0 | 12.3 | 5.0 | $> 100$ | 80 | 0 | 0 |
| Internet peer-to-peer | | | | | | | | | | |
| Gnutella04 | 10876 | 39994 | 996.2 | 96.3 | 16.3 | 6.389 | $> 100$ | 90 | 10 | 2 |
| Gnutella24 | 26518 | 65369 | 997.4 | 97.5 | 17.5 | 7.529 | $> 100$ | 90 | 10 | 4 |

Table 1: Estimates of $\epsilon_t$ for the IC model in several real-world networks: Columns 2 and 3 contain the number of nodes $n$ and edges $|E|$ respectively. There are two sets of measurements of $\epsilon_t$ corresponding to the two methods described in the experiments section. Each set is comprised of 4 values corresponding to different values of transmission probability $p$. In Method 2, the column corresponding to $p = 0.001$ has entries "$> 100$" because it was not possible to estimate the maximum conclusively, as we only considered $\epsilon \leq 100$ in our simulations.

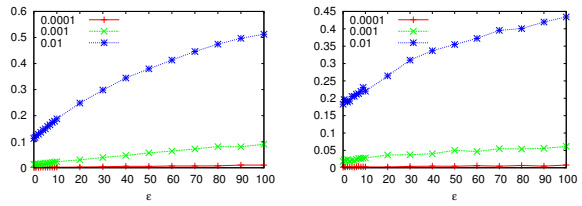
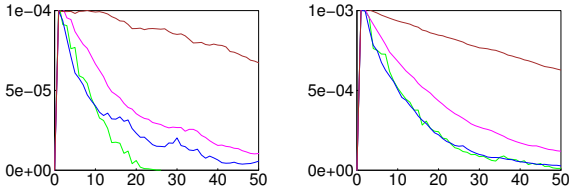



Figure 3: Expected #infections vs $\epsilon$ in the LT model for different seed probabilities $s = 0.0001, 0.001, 0.01$ (seed nodes chosen uniformly at random). Plot (1) Slashdot0811 and (2) Wiki.



(a) s=0.0001 (b) s=0.001

Figure 4: LT model: Plots of infection size over time for Slashdot network for $\epsilon = 0, 1, 10, 100$. Here $s$ corresponds to the seed probability.

## Conclusions and open problems

We give the first rigorous results on the stability of the independent cascades and linear threshold models, with respect to edge perturbations. These help explain our empirical observations on 20 diverse real networks. Extending our results to other models of noise, especially those involving dependencies, sensitivity to the number of sources, and to examine the sensitivity of other dynamical properties in more general diffusion models (including the IC and LT models with heterogeneous probabilities/weights) are natural open problems for future research.

## Acknowledgments

This work has been partially supported by the following grants: DTRA Grant HDTRA1-11-1-0016, DTRA CNIMS Contract HDTRA1-11-D-0016-0010, NSF Career CNS 0845700, NSF ICES CCF-1216000, NSF NETSE Grant CNS-1011769 and DOE DE-SC0003957.

## References

Achlioptas, D.; Clauset, A.; Kempe, D.; and Moore, C. 2009. On the bias of traceroute sampling. *J. ACM* 56(4):21:1–21:28.

Bakshy, E.; Hofman, J.; Mason, W.; and Watts, D. 2011. Everyones an influencer: Quantifying influence on twitter. In *WSDM*.

Borgatti, S.; Carley, K.; and Krackhardt, D. 2006. On the robustness of centrality measures under conditions of imperfect data. *Social Networks* 28:124–136.

Chen, J. 2010. The effects of demographic and spatial

variability on epidemics: A comparison between beijing, delhi and los angeles. In *Conf. on Crit. Inf.*

Chung, F., and Lu, L. 2002. Connected components in random graphs with given expected degree sequences. *Annals of Combinatorics* 6:125–145.

Clauset, A.; Moore, C.; and Newman, M. 2008. Hierarchical structure and the prediction of missing links in networks. *Nature* 453:98–101.

Costenbader, E., and Valente, T. 2003. The stability of centrality measures when networks are sampled. *Social Networks* 25:283–307.

Eubank, S. 2010. Detail in network models of epidemiology: are we there yet? *Journal of Biological Dynamics* 446–455.

Faloutsos, M.; Faloutsos, P.; and Faloutsos, C. 1999. On power-law relationships of the internet topology. In *SIGCOMM*, volume 29, 251–262.

Flaxman, A., and Frieze, A. M. 2004. The diameter of randomly perturbed digraphs and some applications. In *APPROX-RANDOM*, 345–356.

Flaxman, A. 2007. Expansion and lack thereof in randomly perturbed graphs. *Internet Mathematics* 4(2-3):131–147.

Galuba, W. 2010. Outtweeting the twitterers - predicting information cascades in microblogs. In *WOSN*.

Goldenberg, J.; Libai, B.; and Muller, E. 2001. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*.

Gonzlez-Bailn, S.; Borge-Holthoefer, J.; Rivero, A.; and Moreno, Y. 2011. The dynamics of protest recruitment through an online network. *Scientific Reports* 1.

Hagmann, P. 2008. Mapping the structural core of human cerebral cortex. *PLoS Biol*.

Kaneko, K. 1985. Spatiotemporal intermittency in coupled map lattices. *Progress of Theoretical Physics* 74(5):1033–1044.

Kempe, D.; Kleinberg, J. M.; and Tardos, É. 2003. Maximizing the spread of influence through a social network. In *KDD*, 137–146. ACM.

Kempe, D.; Kleinberg, J. M.; and Tardos, É. 2005. Influential nodes in a diffusion model for social networks. In *ICALP*.

Lahiri, M.; Maiya, A. S.; Caceres, R. S.; Habiba; and Berger-Wolf, T. Y. 2008. The impact of structural changes on predictions of diffusion in networks. In *ICDM*, 939–948.

Leskovec, J. 2011. Stanford network analysis project.

Newman, M. 2003. The structure and function of complex networks. *SIAM Review* 45(2):167–256.

Romero, D.; Meeder, B.; and Kleinberg, J. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proc. of WWW*, 695–704. ACM.

Ron, D. 2010. Algorithmic and analysis techniques in property testing. *Foundations and Trends in TCS* 5(2):73205.

Schwab, D. J.; Bruinsma, R. F.; Feldman, J. L.; and Levine, A. J. 2010. Rhythmogenic neuronal networks, emergent leaders, and $k$-cores. *Phys. Rev. E* 82:051911.

Spielman, D. 2009. Smoothed analysis: An attempt to explain the behavior of algorithms in practice. *Communications of the ACM* 76–84.

Wu, C. W. 2005. Synchronization in networks of nonlinear dynamical systems coupled via a directed graph. *Nonlinearity* 18:1057–1064.