Potential Networks, Contagious Communities, and Understanding Social Network Structure

Grant Schoenebeck*
Princeton University
35 Olden St.
Princeton, NJ, USA
gschoene@princeton.edu

ABSTRACT

In this paper we provide evidence that digital social networks look fundamentally different from social networks. We show that online social networks look like a contagion spread over traditional models for social networks. Thus, if these traditional models are correct, then digital social networks and social networks differ in key properties, and we will need different models to describe each. This also indicates that using data from digital social networks may mislead us if we try to use it to directly infer the structure of social networks. Additionally, we describe a framework that we call "potential networks" that may help to use information from digital networks to infer the structure of social networks. Potential networks is a two phase model of social networks. The first phase is the "potential" network. However, this network may not be directly observed and might not even exist an any normal manner. A random process is run over a potential network to produce a behavioral network, the second phase, which can be observed. We then discuss applications of this two-phase framework.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database applications—*Data mining*; G.2.2 [**Discrete Mathematics**]: Graph Theory—*Network problems*; G.3.2 [**Probability and Statistics**]: [Experimental Design]; J.4 [**Social and Behavioral Sciences**]: [Sociology]

General Terms

Theory; Experimentation

Keywords

Social Networks, Community Structure, Random Graphs

*The author thanks the Simons foundation for their generous support of this research.

1. INTRODUCTION

The advent of Web 2.0 has tremendously enriched researchers' access to data. Instead of observing eighteen monks for months waiting for something interesting to happen [23], researchers now have access to approximately 160 millions users' 90 million daily tweets through Twitter's API [1]. While looking at the data clearly tells us a lot about what people do online, it is less clear how much this data tells us about people. Social science researchers developed social networks as a methodological tool for understanding social phenomena, such as how individuals' actions affect macro-level features of society, or how an individual's "location" in a network affects his/her opportunities. Social networks are not to be conflated with online social networks such as LiveJournal, Epinions, MySpace, Facebook, and Twitter. We will use the term *digital networks* to include online social networks as well as other digitally recorded networks, e.g. citation networks, communications networks, collaboration networks, web graphs, co-authorship networks, product co-purchasing networks. Digital networks provide a means for studying social questions pertaining to social networks by providing a lot of data. However, it is difficult to know how close this data generalizes past the digital world. To make full use of this data to make assertions about social questions, we need to know that it does generalize, at least in the cases that we care about.

Sociologists have long distinguished between different types of networks. One example is trust networks: from whom would you feel comfortable asking for $1000? Another example is self-declared, or articulated, networks: who do you want the world to believe are your friends? While differences such as these are sometimes ignored in the greater social networking community, many researchers still hold on to their importance as indicated by a familiar question, "who in the room is friends with his/her mother on Facebook?" However, even if no one were Facebook friends with his/her mother would this really affect any large-scale measurements of the data? Does the shear scale of such data render differences between the digital social networks and social networks to be mere annoyances or do these differences present a substantial obstacle to using data from digital social networks to make inferences about social networks. This is a question that this paper hopes to address.

To put it another way: Is analyzing digital social networks tantamount to holding up a big mirror to our society, or is more like looking at our society in a fun-house mirror–

where things may appear much different than they are–you may only see feet and a head? And if digital social network data is akin to looking in a fun-house mirror, then what aspects of reality can we still reliably deduce from looking at this data?

There is a long line of work that seeks to study network generation models (for examples see [8, 4, 27, 3, 18, 17]). Such studies often observe a property in many different networks and show that the current models fail to exhibit this property. They then propose a new model that does exhibit this property. This work is implicitly based on the assumption that all networks share certain common properties, and thus can be accurately modeled by some random generative graph model. Indeed there are universal properties that span a large array of sample networks including local clustering with short average distances [25] and long tail degree distribution [3]. However, we should not *a priori* expect all properties to be universal, and thus we should not expect one generative model. One of the original motivations for sociologists to develop social network theory was to explain how these networks *differ* and the implications of these differences. For example Gans [9] studied how Boston's West End community was unable form a coalition to fight a "reorganization" measure that ended up destroying the community, even though other seemingly similar communities were able to organize against and defeat such measures [12].

In this paper we provide evidence that digital networks look fundamentally different from social networks. Recent work has shown that digital networks contain properties not present in many traditional generative graph models [17, 18]. On the one hand, if these traditional models are correct, then digital networks and social networks differ in key properties, and we may need different models to describe each. On the other hand, traditional models could just be flawed with respect to these properties and need to be remedied. We provide evidence of the former. We create a model of digital networks being created similarly to a contagion spreading over a existing social network. This model is simple and natural and allows us to use a traditional generative model (the Watts-Strogatz model) and yet produce digital networks that contain properties observed in a digital network that *are not* contained in the staring model. If indeed social networks and digital networks are different, this indicates that using data from digital social networks may mislead us if we try to directly use it to understand how networks form. It further raises methodological questions about mining large data sets for properties to directly attribute to social network models, and then coming up with models that have these properties. (However, it does allow for the mining of large data sets for properties to directly attribute to *digital network* models.)

Additionally, we describe a framework that we call "potential networks" that may help to use information from digital networks to infer the structure of social networks. Potential networks is a two phase model of social networks. The first phase is the "potential" network. This network may not be directly observed or even exist an any normal manner. The second phase is the "behavioral" network, which is observable. However, the behavior network is realized by running some random process over the potential network to produce the behavioral network.

We will expand on the definition of a potential networks later in the paper, but here we define them in a simple, restricted manner (which we later call the *static* model). Let UG denote the set of undirected graphs. Then a *Potential/Behavioral Network* denoted $PBN(G, D)$ is a simply an undirected graph $G = (V, E) \in UG$ and a dynamics $D : UG \to UG$ which is a possibly randomized function from the set of undirected graphs to itself in such a way that $D(G)$ is an subgraph of G. The idea is that G is a list of "potential" nodes and "potential" edges and $D(G)$ is a list of "observed nodes" and "observed" edges.

Road Map. In Sectionřef'Contagious Communities' we argue that digital networks can be naturally modeled by spreading a virus over a typical social network model. We run simulated test of these models and claim that the results both fit intuitions and empirical data about social and digital networks. In Section 2.3 we draw implications of this model. In section 3 we construct a framework to indirectly infer social networks using data from digital network. We show how many of the past results and some open problems fit into this framework. In Section 3.3 we briefly discuss additional related work. Finally, we conclude with a summary of main points in Section 4.

2. CONTAGIOUS COMMUNITIES

In this section we will illustrate a case where the "behavioral" network exhibits very different large-scale characteristics than the "potential" network. In this example, the potential network is the extant social network, and the behavioral network is an online social network (e.g. LiveJournal[1]).

Data mining has shown that digital social networks all share a few common features: Power-law degree distributions, shrinking diameters, and a particular "network community profile plot". We show, with computer simulations, that even though the Watts-Strogatz model has none of these properties, if we use it as a potential network and spread a contagion in a natural way that the resulting network has all of these properties.

Power-Law degree distributions. Previous research has shown that many of the degree distributions are power-law distributions[3]. This means that when the degree distribution is a straight line when plotted with both axes logarithmically scaled. Often this serves as a contrast to Poisson distributions, which are much more highly concentrated and have a much thinner tail (fewer points far from the average).

We use the term power-law in a very loose sense: by power law, we will mean that the log-log plot of the degree distribution of the nodes of the graph appears roughly linear, nothing more. While this is not the true meaning of the term, it does capture the operation definition of the term in many other papers and so we use it as well (see discussion on page 60 of [13]). Those uncomfortable with this use can

[1]LiveJournal is an early blogging and social networking community. We will use it as a running example for a digital social network.

substitute the term skewed distribution instead of power law distribution.

Shrinking Diameters. Previous research has also shown that, over time, the diameter of digital networks tends to shrink [18]. This work was based on analyzing four networks: the ArXiv citation graphs (for high-energy physics theory), the U.S. Patent citation graph, the graph of routers of the Internet, and the ArXiv affiliation graph (on certain topics). Note, however, that none of these is actually a digital social network.

Network Community Profile. Another network feature that we are interested in is called the *network community profile* and was described in the paper "Statistical Properties of Community Structure in Large Social and Information Networks" by Leskovec, Lang, Dasgupta, and Mahoney [20]. In this paper, the authors develop a new way to analyze a network that they call the "network community profile"-which we will describe shortly. They then show that this tool looks similar on over 70 real data sets that they have access to, such as LiveJournal. In particular, the network community profile plot on the social networks: LiveJournal, Epinions, LinkedIn, Del.icio.us, and Flickr look nearly identical (see [20] pages 22 and 25). They note that the plot decreases until around 100, then it stays roughly even for a short period, and finally starts to increase. Finally, they show that this tool looks completely different on virtually all generative models (except for one that they call the Forest-Fire model). This is pretty shocking.

Leskovec et al were interested in studying the community structure on networks. They define a community as a set of nodes with low conductance (i.e with many edges within the set compared to the number of edges leaving the set). Even in very large datasets of digital networks, they found few large communities (over 100 people) that fit this definition. Broadly speaking, they found that the structure of these graphs was composed of "whiskers" and a "core". *Whiskers* are a set of nodes connected to the rest of the graph by only a one or a few edges. The *core* was a big connected mess with no subsets of small conductance. The "community" structure that they detected (sets with low conductance) could be almost entirely attributed to collections of whiskers–groups just barely connected to the rest of the graph.

However, despite this result, we will show that this does not mean that we need to throw away all our models just yet. It may be that these over 70 data sets all have attributes of digital networks which are not shared by social networks. Thus it could be that the generative models do well simulating social networks, but poorly simulating digital networks. On the positive side, this would mean that not all our previous models are useless. On the negative side, it may be impossible to find a perfect generative model because digital networks look different than social networks.

The intuition behind our model is that digital networks (e.g. LiveJournal) *spread* over extant social networks, so that people join an online social network because one or more of their friends already participate in that social network. Thus to model online social networks, we should start with a model of a social network and then model a process of individuals joining the online network.

We simulate exactly this. We start by creating a potential network using a simple generative model–the Watts-Strogatz model, which does not exhibit community network profiles that match the 72 data sets as a potential network. We then generate a behavioral network based on spreading a contact process on this potential network and look at the resulting subgraph of infected nodes. We show that this does exhibit the matching community network profile. Again, the observation that this behavioral model is based on is that people join networks because their friends are on them. Intuitively, this is true across is large family of networks: from joining LiveJournal to authoring an astrophysics publication networks, from acting in a movie to using Twitter.

2.1 The Model

Watts-Strogatz model. The Watts-Strogatz random network model is define by three parameters. The undirected $WS(n, d, r)$ ensemble of random graphs–where n is the number of vertices, d is the average degree, and $r \in [0, 1]$ is a parameter–is defined by the random process that creates them. This process begins with the graph on n nodes $\{0, 1, \ldots, n-1\}$ where each node is connected to the d closest other nodes so that $E = \{(k, k \pm \ell) : 1 \leq \ell \leq d/2\}$. Each edge (u, v) is then "rewired" with probability r, that is replaced with the edge (u, v') where v' is chosen from the vertices not already connected to u.[2]

Model of Transmission. In this section we first define two simple models of transmission.

The first we call *random edge transmission induced graph* which has one parameter. $RATIG_G(m)$ is defined by starting with the graph $G = (V_G, E_G)$ and initiating the infected set I to a singular random vertex. A random edge (u, v) is chosen uniformly from $E(I, \bar{I})$ and the vertex v is added to I. This is repeated until $|I| = m$. The resulting infected graph is $G(I)$, the induced subgraph of G on the vertices in I.

The first model includes all the edges in G between vertices that are in I. In the second model, these edges must also be discovered. We call the second model *random edge transmission*, and it has three parameters. $RET_G(m, \alpha, \beta)$ is defined by initiating the infected graph $H = (V_H, E_H)$ to the graph $(\{v_0\}, \emptyset)$ where v_0 is a random vertex from the potential graph $G = (V_G, E_G)$. At each step, each edge $(u, v) \in E_G(V_H, V_H) - E_H$ is added to H with probability α and each edge $(u, v) \in E_G(V_H, \bar{V_H})$ is added to H (along with v) with probability β. The process is run until m additional vertices are included.

Note that $(G, RETIG_G(m))$ and $(G, RET_G(m, \alpha, \beta))$ are Potential/Behavioral Networks.

[2]It is a little more complicated than this because the order which you consider the edges may matter, see [27] for the details of ordering.

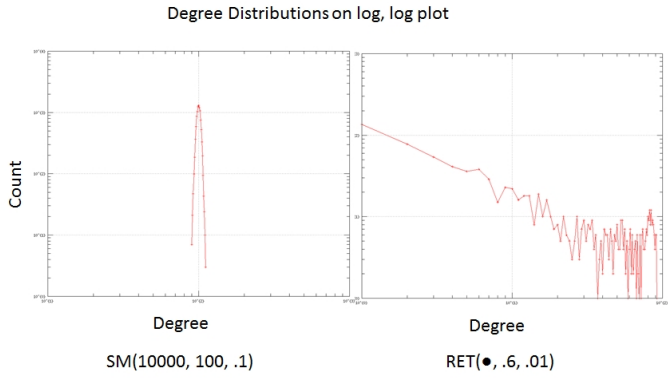

Figure 1: A degree plot of $WS(1000, 100, .1)$ and $RET_{WS(10000,100,.1)}(1000, .6, .01)$

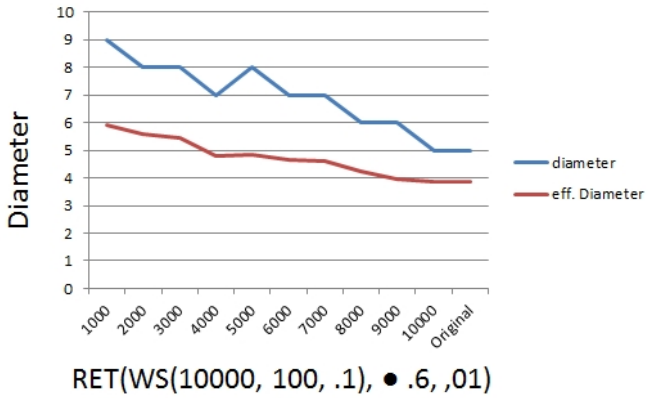

Figure 2: A plot of the diameter and effective diameter as a RET contagion was spread on $WS(10000, 100, .1)$ with $\alpha = .6$, and $\beta = 0.01$

We create a more complex model which allows people in "infected" communities to make new friends.

The *random edge transmission with exploration* which has four parameters: $RETWE_G(m, \alpha, \beta, \gamma)$ is defined exactly like the *random edge transmission induced graph* except that at each round, for each triple $u, w, v \in V_H$ where $(u, w), (w, v) \in E_H$ the edge (u, v) is added to E_H with probability γ (this edge is added with probability γ for each such triple).

Note that as we defined potential networks above, this is *not* a potential network, but it will turn out to be a dynamic potential network.

Network Community Profile Plot. The network community profile plot[19] is based upon the idea of the conductance. The *conductance* of a set $S \subseteq V$

$$\Phi(S) = \frac{E(S, \bar{S})}{min\{degree(S), degree(\bar{S})\}}$$

is equal to the number of edges leaving a set S divided by the sum of the degree of the vertices in S (or $\bar{S}$, whichever is smaller). Thus, if S is insular and does not have many edges leaving it relative to its total degree, then S has low conductance. The community network profile finds the set of each size $s : 1 \leq s \leq |V|/2$ with the lowest conductance and then plots this graph. I.e $f_G(x) = \min_{S:|S|=x} \Phi(S)$.

2.2 Simulation Results

In this section we describe the results of our simulations. All simulations were done using the SNAP System [16]. This is particularly important for the network community profile plots which, because the it is NP-complete to computer exactly, is approximated with with heuristics. These heuristics were shown to work well in other graphs (see Section 5 of Leskovec et al [20]), but that is no guarantee that they work well here.

We first explain what happens when we use the Watts-Strogatz model as a potential graph. We study three properties of the network: degree distribution, diameter, and network community profile.

Degree Distribution. Interestingly, even though the degree distribution of the Watts-Strogratz model is extremely concentrated, we found that the resulting graphs from RET had a degree distribution that resembled a power-law distribution. In particular, we found that if took the resulting graph from $RET_{WS(10000,100,.1)}(1000, .6, .01)$ that the degree distribution appeared to be a power-law distribution[3]. Recall the $WS(n, d, r)$ is the Watts-Strogatz model of graphs with n nodes, average degree d, and rewiring probability r, and that $RET_G(m, \alpha, \beta)$ is the contagion model on graph G where at each round, edge between infected nodes is added with probability α and each edge from an infected node to a non-infected node is added with probability β until there are m nodes that are infected.

This is especially surprising since the maximum degree of the original graph G (and hence the largest possible degree in H was only a little over 100 in the trials we ran. We did not expect to find this, and we illustrate the difference Figure 1 Of course, as the infected graph becomes a large fraction of the network, we expect this effect to go away. In fact, the effect seemed to disappear when the infected region reached 40% of the vertices, and disappeared as soon as 20% in some trials.

Diameter. We also observed the diameter of the network. We found the the diameter shrunk in according with the prediction of Leskovec, Kleinberg, and Faloutsos [18]. We plot the diameter and effective diameter of one of the runs in Figure 2.2.

Network Community Profile. We found that if took the resulting graph from $RET_{WS(10000,100,.1)}(1000, .7, .01)$ that the network community profile closely matches that of the online social networks that Leskovec et al studied in [20]. Recall the $WS(n, d, r)$ is the Watts-Strogatz model of graphs with n nodes, average degree d, and rewiring probability r,

[3]see beginning of this Section for our relaxed definition of a power law distribution.

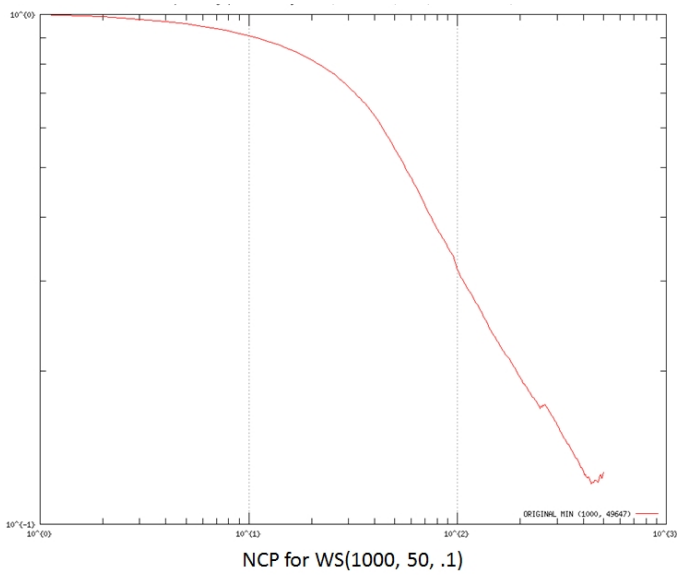

Figure 3: A ncp of WS(1000, 50, .1)

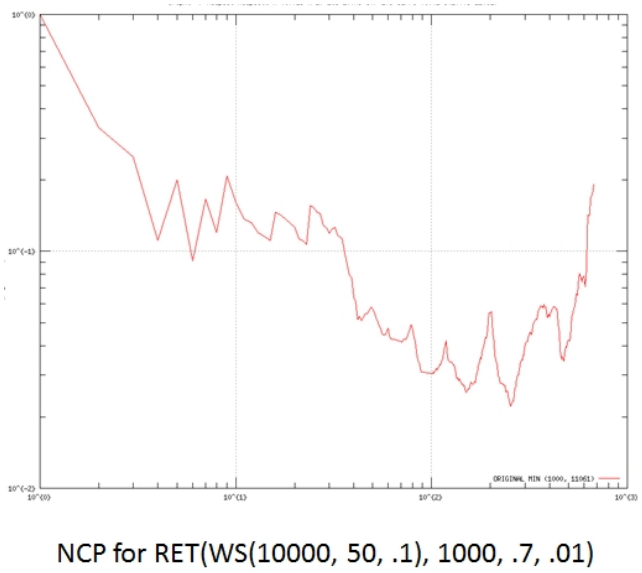

Figure 4: A ncp of RET run on WS(10000, 50, .1) with $\alpha = .7$, $\beta = 0.01$, and $m = 1000$

and that $RET_G(m, \alpha, \beta)$ is the contagion model on graph G where at each round, edge between infected nodes is added with probability α and each edge from an infected node to a non-infected node is added with probability β until there are m nodes that are infected. Figure 1 and Figure 2 respectively show both the original network community profile of the Watts-Strogatz model and the network community profile of a virus spread over the network.

This result was robust to changing the size of the network n and the degree of the network d. However, if we made r too large, then the community network profile would look different, the plot never decreases sufficiently. A similar pattern would happen if α were not sufficiently large compared to β, the edges between nodes of the infected graph H would fail to fill in and no community structure would be detected.

Finally, as the size of the infected graph H approached the entire graph G, then the community network profile of the infected graph H would look increasingly like the community network profile of G. This is because eventually H will become G.

Other Graph Generation Models. When we run this process on various graph generation models including Erdos Renyi random graphs [8], Preferential Attachment networks [3], or complete graphs we did not see this behavior. We hypothesized that, in the Erdos Renyi random graphs and the Preferential Attachment model, this is because there very little clustering to begin with, and the virus spread evenly over the graph in a tree like fashion and remained unclustered. Such behavior might not continue if nodes "met" other infected nodes by virtue of being infected and having a common neighbor. To test this hypothesis we embellish the dynamics to artificially add community structure using $RETWE$ as a model of spreading. We find that while the network community plot looks more like the typical plot, it still does not look like the sought after behavior.

The fixing of parameters are robust. We emphasize that the results here are not cherry picked; we reported everything we tried that did not work, as well as that which did. It should also be pointed out that the Watts-Strogatz model is simply the first model that we tried this on which worked. We do not claim that it is the only model over which these results of contagious communities will hold.

Forest Fire Model as a Contagious Community. Leskovec et al found that the Forest Fire Model was the one model they tested that did replicate the results of the community network profile that they found on the 72 data sets. We note that the Forest Fire Model can actually be reinterpreted as a potential behavioral network model.

The complete Forest Fire model can be found on page 9 of Leskovec, Kleinberg, and Faloutsos [18]. For our purposes, it will suffice to present a slightly simplified undirected version. Our Forest Fire model has one parameters p, the burning probability. The model starts with a single node. At each time step a new node v joins and chooses an existing node u at random and forms a link with u. For each node w that v links to (starting with u), v also links to k_w of w's neighbors where k_w is chosen from a binomial distribution with mean $(1-p)^{-1}$. This is guaranteed to terminate because v is not allowed to link to any node more than once.

The Forest Fire model is very close to running $RETWE$ on a low degree random Erdos-Renyi graph. Consider running $RETWE$ on a low degree random. When a vertex v joins (if the graph is not more than a small fraction of the total), then it is very likely that v is attached to exactly one node u of the infected subgraph, H, (the one the infected v). Thus the principal way for v to add more ties in the infected subgraph

H is by exploration on the infected subgraph through ties of u in H. Each time that v links to a neighbor w of u in H, the next time step v can add nodes to neighbors of u as well (v can also add ties by infecting a neighbor in G that is not yet in H). The difference between these two models is in the number of neighbors that u finds by exploration. In the forest fire model it is $(1-p)^{-1}$ in expectation, and in the $RETWE$ model it depends on the amount of time the node has been in the network.

Thus it is not surprising that both of these models produce similar though not certainly not identical network community profile plots.

2.3 Conclusions on Simulations

We think that there are several important lessons from the above simulations.

Digital and social networks may not look the same: We show that metrics that appear to test global properties (e.g network community profile) and metrics that appear to test local properties (e.g. degree distribution) may show dramatically different results on digital networks and social networks. While people have made this observation before, here we provide results that begin to show the scope and scale of the qualitative and quantitative differences. Moreover, our two step simulation captures both the intuition and sociology research about social network models–small diameter [21] and local clustering [25]–and the research on digital social network–shrinking diameter [18], power law degree distribution [3], and network community profile plot [17]–all in one simple and intuitive model. We show that if the intuition that guided the first generative models is correct, then this discrepancy must exists. It could still be that the core and whisker model characterizes actual social networks. However, we feel that contagious communities is a more natural explanation for the whisker and core structure observed on digital networks.

This indicates that we may not be able to learn about the structures of social networks by directly data-mining digital networks. This is a reality check on some of the work done which is motivated, in part, by the quest for a universal generative model.

Digital networks are not as we would commonly believe them to be: Leskovec et al, already showed us that *digital networks are not as we would commonly believe them to be.* This shows us that this conclusion may not reach to other social networks. Some doubted the counter-intuitive results of Leskovec et al. Our experiment supports their work, but only in the their original scope of digital networks. Similarly, this provides intuition behind the observations of Leskovec et al of the odd properties that all the digital networks they considered seemed to have.

New generative model for digital social networks: This intuition also provides us with a *new generative model for digital social networks.* Start with a social network model, and model a contagion spreading over it.

Re-imagine what a community is and what they look like: Secondly, and more speculatively, this gives us an opportunity to *re-imagine what a community is and what they look like.* These digital social networks can be seen as a *community* within the potential network. That is the nodes of LiveJournal form the "LiveJournal community" which is embedded in society. The LiveJournal network can be viewed both as a network in and of itself, but also as a community in a larger network. We can perhaps use the core/whisker model of Leskovec et al to understand properties of a community.

This type of community is much different than the traditional community that researchers have looked for–namely a set of nodes with many internal but few external edges. Such a community is an *insular* community. The notion of insular community can be captured by such metrics as modularity [22], and conductance [19]. However, this model provides an alternate view of what a community looks like–communities that are gradually adding internal connections and external members that start on the periphery of the group and gradually become more central to the community. Communities are composed of a core that has no insular communities and whiskers.

Digital social networks have boundaries: This view of digital networks as communities created in a larger social network also allows us to see that *digital networks have boundaries.* Difficulties of boundary definitions often arise in studying social networks. Researchers much decide on the exact scope of the community which they are studying. Subtle differences on where exactly boundaries are placed can affect metrics of the community such as average degree, average distance, et cetera [11]. While the LiveJournal graph may seem to be void of this difficulty, it is not. The LiveJournal community exists in time as well. If you look at the nodes and edges at a particular moment in time, you may see the "whiskers" that Leskovec et al observed. However, these almost disconnected communities may eventually meld in with the core, but simply have not yet. By only looking at the edges before a particular time, the network may contain features, such as groups of vertices barely connected to the main part of the graph, that are not present in a more holistic view of the network.

Potential Networks: Lastly, these results point toward the potential network frame-work developed in the remainder of the paper. A model were social networks are not created *ex nihilo*, but from existing social structures.

3. POTENTIAL NETWORKS

We start by defining a Potential/Behavioral Network Model. We define two such models, a static model and a dynamic model. We first define a network, which will be a graph with additional information on the edges and vertices.

Definition 1. A *Network Model* $\mathcal{N}$ is a pair $\mathcal{N} = (\Omega_V, \Omega_E)$ where Ω_V and Ω_E are each sets. Ω_V is a set of possible vertex attributes and Ω_E is a set of possible edge attributes. A realization of a network model $\mathcal{N}$ is a network $N = (G, A_V, A_E)$ where $G = (V, E)$ is an undirected graph, $A_V : V \to \Omega_V$, and $A_E : E \to \Omega_E$ assign attributes to each vertex and edge of the graph. We will denote by $G(N)$ the undirected graph associated with the network N.

Undirected graphs, directed graphs, weighted graphs, graphs with different edge types, graphs where each edge exists only in certain intervals of time, etc are all network models.

Definition 2. A *Static Potential/Behavioral Network Model* is a triple $(\mathcal{M}, \mathcal{N}, D)$ where $\mathcal{M}$ and $\mathcal{N}$ are network models, and $D : \mathcal{M} \to \mathcal{N}$ is a (probabilistic) function from a network $M \in \mathcal{M}$, called the potential network, to a network $N \in \mathcal{N}$, call the behavioral network, such that $G(D(M))$ is a subgraph of $G(M)$.

A *Potential Network* for a model $(\mathcal{M}, \mathcal{N}, D)$ is simply a network $M \in \mathcal{M}$

A *Dynamic Potential/Behavioral Network Model* is a triple $(\mathcal{M}, \mathcal{N}, D)$ where $\mathcal{M}$, and $\mathcal{N}$ are network models, and $D : \mathcal{M} \to \mathcal{M} \times \mathcal{N}$ is a (probabilistic) function from a network $M \in \mathcal{M}$, called the potential network, to a network $M' \in \mathcal{M}$ called the updated potential network, and $N \in \mathcal{N}$, call the behavioral network, such that $G(D_2(M))$ is a subgraph of $G(M)$, where $D_2(M)$ denotes the second component of $D(M)$. This definition can by applied iteratively over time so that $M_0 = M$ and $(M_t, N_t) = D(M_{t-1})$ producing a sort of Hidden Markov model.

The static potential/behavioral network model is conceptually simpler. However, it ignores the fact that current outcomes can affect the feasibility of future outcomes. In the dynamic model, the way that the behavioral network is realized at time t can actually change the potential network at future steps. This models the fact that if two people meet, they may be more likely to meet up in the future. Additionally, it captures the notion that making ties in the present can enable the creation of future ties that are not currently possible.

We observe that many common network generation models can naturally be seen as potential networks.

We start with defining Erdos-Renyi graphs this way. We first describe the potential/behavioral network model, which we call the *Basic Model*.

Definition 3. The *Basic Model* is the Potential/Behavioral Network Model $(\mathcal{M}, \mathcal{N}, D)$ where $\mathcal{M} = (\{1\}, [0, 1])$ (weighted undirected graphs), $\mathcal{N} = (\{1\}, \{1\})$ (undirected graphs), the dynamics $D(M)$ applied to network $M = (G, A_V, A_E)$ creates a graph with the same vertices of M such that each edge $e \in G(M)$ is present with probability $A_E(e)$ (the edge weight in M).

The Erdos-Renyi graphs then simply running the Basic Model starting with the potential network $(G = (V, V \times V), \vec{1}, \vec{p})$ so that G is the complete graph, for each vertex $A_V = 1$, and for each edge $A_E = p$.

In fact, the Basic Model can simulate several additional models. One is the *planted community model*. The planted community model has four parameters (n, k, p, q) where $k \leq n$ are integers such that k divides n, and $p \geq q \in [0, 1]$. The graph is the complete graph on n vertices. The vertices are partitioned evenly amongst the k groups, and the weights of edge (u, v) is p if v and u are in the same group, and q otherwise.

Another model that can be simulated is the *planted clique model*, which has two parameters k and n. The potential graph is again the complete graph G on n vertices. k special vertices are randomly chosen to be in the planted clique. The edge weights are 1 for edges that connect two special vertices, and are 1/2 for all other edges.

Additionally, the Small World model of Watts and Strogatz [27], the and the navigational small world models of Kleinberg [15] and Watts et al [26] can be captured by this a slight variant of the Basic Model.[4]

Finally, as pointed out in Section 2.2, the Forest Fire Model [18] can be simulated by $RETWE$, which is a natural dynamic potential/behavioral model. Each time a link (u, v) is added in H, u gains access to all of v's numbers in H, even if v does not have access to them in G. This is a dynamic potential/behavioral network model because the potential network changes to reflect these new possibilities.

3.1 Using Behavioral Networks to Recover the Potential Graph

While in Section 2 we showed the we may not be able to directly recover the potential graph from a behavioral graph, not all hope is lost. We may be able to recover the potential network indirectly, or at least to recover properties of the potential network. This is one of the key motivations of the definition of Potential/Behavioral networks. Many open questions can be cast in this way, and we discuss some of them now.

Even if the graph was not produced using a potential network, the behavioral/potential framework can still be used to learn about aspects of the graph. For example, in the planted community model of the previous section, the goal is to recover the initial communities given the behavioral graph. Given a graph H on n vertices, if we assume that it was created by running the planted community model with parameters n, k, p, q with some fixed k and any $0 < q < p < 1$, then finding a maximum likelihood partition of the vertices (in the planted community model) will give a partition of the graph into k equally sized components that minimized the number of edges crossing the components. When $k = 2$ this is the balanced separator problem which is NP-hard to even approximate [14], so we cannot hope to solve it in the worst case. However, we can hope to solve it in the average case and in fact even rather simple heuristics work for the case of the planted community model. [5].

Some properties that we may want to reconstruct are average degree, finding high influence nodes, testing for hierarchical structure, testing if a network is resilient to failure, finding communities, finding structure holes and more. The amount, accuracy, and type of data required for testing each

[4]The Basic Model simulates $G_{n,p}$ random graphs where as these models are really random graphs with fixed degree, and so are simulated by a slightly different model that can condition on particular degrees.

of these properties is likely to differ.

We can also use this model to derive *confidence intervals for graph properties*. An overall strategy to recover properties of the potential graph H is to assume $H = D(_{R\to}G)$ where G is produced from some generative model R and D is some dynamics. In this way we can provide average case solutions to problems with worst-case hardness. An key application of this model is when D is some social network sampling procedure. We would then like to know what kind of conclusions we can safely draw by only assuming a generative model R. This is similar to linear regressions, where it is assumed that the data comes from a linear model (R) with error (D) and the goal is to recover the model by eliminating the error, and also to provide a confidence of this model. Here we would like to do the same, but many graph properties are very non-linear and the sampling error D can be correlated with the structure G that we are trying to identify. We would like to know what kinds of error we can tolerate, and what kinds of error blow up in a non-linear fashion.

Results for this potential/behavioral model could prove useful for social network research methodology because of the non-linear nature of many graph properties, existing models often fail to capture the complexities of the setting and using such models renders the confidence intervals useless [11].

A prerequisite for using such a model is having a good model for D. In the case of contagious communities D is a local property, and so may be easier to estimate than a more global graph property. In the contagious communities model it is important to estimate the parameters α, β, and γ appropriately, without such estimates it is unclear what can be learned. It seems like the best way to estimate such parameters may be to understand what is happening offline (at least in restricted settings). For example, do users of LiveJournal meet new people on the site, or simply connect in new ways with people they already know? While we may try to infer it from network data, another possibility is to use studies that involve directly observing people to estimate reasonable models of dynamics.

3.2 Subtleties

It is sometimes important to distinguish between potential networks and behavioral networks, but other times this distinction seems insignificant. An added benefit of carefully defining this framework is that we can clear up some ambiguities.

A case where it is important is in the hidden clique model defined above. In this model, it is unknown (and conjectured to be hard) to find the large hidden clique in $D(m)$ even when $k = o(\sqrt{n})$ [2]. However, note that the clique is trivial to find in the potential graph G itself! Additionally, if the dynamics D change so that the resulting graph is not the result of one sampling, but the result of $2\log(|V|)$ samplings, then again the problem is easy[5] Even if the clique is replaced by a dense subgraph, the problem remains easy with $\Omega(\log(n))$ samples (by a standard Chernoff bound)[6]

[5]Simply discard all edges that do not appear in every sample, and with high probability the only remaining edges left will be those of the clique.

[6]With high probability the only edges that will appear in at least a $(p+q)/2$ fraction of the instances are those of the clique.

This means that if we do not only have data on which edges are present, but only, say, who talked to whom over the last year, we can efficiently compute a lot more than with a sampled graph alone.

A case where distinguishing is not important is in a simple contact process (e.g. SIR model). The distribution of graphs where each edge is present with probability p and a contagion spreads across an edge with probability q, is the same as a model where the graph is complete and each edge has weight p, and the contagion spreads with probability qw where w is the weight of the edge.

3.3 Related Work

The Potential/Behavioral Network framework has implicitly been used in many previous studies. Segal [24] observed that the best prediction of who would become friends at a certain police academy was the proximity of their last names in the alphabet (this was presumably due to the frequent placement of the cadets in alphabetical order). Thus the last names produced a certain potential network. In another study, also at a police academy, Conti and Doreian [7] show that seating assignments and squad assignments heavily predict friendship ties. However, they also showed that this effect lessened slightly with time. These studies can be reinterpreted as trying to understand the underlying potential networks. In the later case, they wanted to manipulate these networks to foster inter-racial comradery at the academy.

In an experimental study, Centola [6] created digital communities populated with volunteers and studied the spread of joining health forum network over this community. Thus Centola was studying how the current network influences individuals to join a website (and thus a new network). Viewed in the potential/behavioral framework, this is clearly a contagious community with a strictly enforce potential network). Centola was mostly concerned with what types of potential networks would foster the largest contagion.

Additionally, recent work by Gomez-Rodriguez, Leskovec, and Krause [10] creates a model to try to infer a network of influence by looking only at the time sequence of an infectious outbreak (e.g. a news item through the blogosphere). They show via computer simulations that their heuristics for recovering a potential network, given the timing data from a series of outbreaks, can simultaneously give high precision and recall of the original edges.

4. CONCLUSIONS

In this paper we have provided evidence that digital and social networks differ in foundational properties. We did so by creating a new network generation model that spreads a contagion across an existing network. We show that empirically, this model has a realistic network community profile, power law distributions, and shrinking diameter, yet also conforms to traditional generative models. Moreover this model allows us to re-imagine what a community is, and what it looks like on a social network–perhaps more like a core and whisker model than an insular community. Also this model

shows that even with an entire digital social network, there are still boundary problems to contend with.

Finally, we provided a theoretical framework that is broad enough to accommodate many open question. Our framework shows how certain computationally intractable problems can become tractable with slight changes in the model. Also, this framework provides a way of creating new methodological tools in social network analysis.

5. ACKNOWLEDGMENTS

I would like to thank the many people whose help and insights have influenced this paper including Sarita Yardi, danah boyd, Michael Mohony, and Yuri Leskovec. Also, I would like state my appreciation for SNAP the Stanford Network Analysis Platform.

6. REFERENCES

[1] Twitter about page, October 2010. `http://www.twitter.com/about`.

[2] N. Alon, M. Krivelevich, and B. Sudakov. Finding a large hidden clique in a random graph. In *SODA '98: Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms*, pages 594–598, 1998.

[3] A. Barabasi and R. Albert. Cemergence of scaling in random networks. *Science*, (286):509–512, 1999.

[4] E. A. Bender and E. R. Canfield. The asymptotic number of labeled graphs with given degree sequences. *Journal of Combinatorial Theory, Series A*, 24(3):296 – 307, 1978.

[5] T. Carson and R. Impagliazzo. Hill-climbing finds random planted bisections. In *SODA '01: Proceedings of the twelfth annual ACM-SIAM symposium on Discrete algorithms*, pages 903–909, Philadelphia, PA, USA, 2001. Society for Industrial and Applied Mathematics.

[6] D. Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 2010.

[7] N. Conti and P. Doreian. Social network engineering and race in a police academy: A longitudinal analysis. *Social Networks*, 32(1):30 – 43, 2010. Dynamics of Social Networks.

[8] P. Erdos and A. Renyi. On the evolution of random graphs. In *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, pages 17–61, 1960.

[9] H. J. Gans. *The urban villagers : group and class in the life of Italian-Americans / by Herbert J. Gans ; foreword by Erich Lindemann.* Free Press ; Collier-Macmillan, New York : London :, 1962.

[10] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *KDD '10: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1019–1028, 2010.

[11] R. Grannis. Six degrees of "who cares"? *American Journal of Sociology*, 115(4):991–1017, 2010.

[12] M. Granovetter. The Strength of Weak Ties. *The American Journal of Sociology*, 78(6):1360–1380, 1973.

[13] M. O. Jackson. *Social and Economic Networks.* Princeton University Press, 2008.

[14] S. Khot and N. Vishnoi. The unique games conjecture, integrality gap for cut problems and the embeddability of negative type metrics into ℓ_1. In *Proceedings of the 46th IEEE Symposium on Foundations of Computer Science*, pages 53–63, 2005.

[15] J. Kleinberg. The small-world phenomenon: an algorithm perspective. In *STOC*, pages 163 – 170, 2000.

[16] J. Lescovec. Snap: Stanford network analysis program, 2010.

[17] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, New York, NY, USA, 2008. ACM.

[18] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187, New York, NY, USA, 2005. ACM.

[19] J. Leskovec, K. Lang, A. Dasgupta, and M. Mahoney. Statistical properties of community structure in large social and information networks. In *Proceedings of the 17th International World Wide Web Conference*, 2008.

[20] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1):29–123, 2008.

[21] S. Milgram. The small world problemt. *Psychology Today*, 1:62–67, 1967.

[22] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69(2):026113, 2004.

[23] S. F. Sampson. *Crisis in a cloister.* PhD thesis, Cornell University, 1969.

[24] M. Segal. Alphabet and attraction: An unobtrusive measure of the effect of propinquity in a field setting. *Journal of Personality and Social Psychology*, 30(5):654–657, 1974.

[25] D. Watts. Networks, dynamics, and the small world phenomenon. *American Journal of Sociology*, 105(2):493–527, 1999.

[26] D. Watts, P. Dodds, and M. Newman. Identity and search in social networks. *Science*, 296:1302–1305, 2002.

[27] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, (393):440–442, 1998.