Chapter 11
Propagation Models and Analysis for Mobile Phone Data Analytics

Derek Doran and Veena Mendiratta

Abstract People in modern society use mobile phones as their primary way to retrieve information and to connect with others across the globe. The kinds of connections these devices support give rise to networks at many levels, from those among devices connected by near-field radio or bluetooth, to society-wide networks of phone calls made between individuals. This chapter introduces state-of-the-art propagation models that have been applied to understand such networks. It discusses how the models are used in many innovative studies, including how short-lived information spreads between phone callers, how malware spreads within public places, how to detect fraudulent and scamming activity on a phone network, and to predict the propensity of a user to unsubscribe from a mobile phone carrier. It concludes with a discussion of future research opportunities for the study of propagation modeling to mobile phone data analytics.

11.1 Introduction and Motivation

As of February 2013, an astonishing 6.8 billion mobile phone subscriptions are active across the world.[1] This huge number of subscribers, constituting a majority of the world's population, reflects how citizens of countries with varying socioeconomic conditions all rely on cellular devices to communicate and connect with others. These devices, which are typically full of data about who our contacts are, the kind of information we share, who we communicate with, and our physical location have also emerged as an attractive platform to study human behaviors and activity across large geographic regions. For example, the analysis of mobile phone data has

[1] http://mobithinking.com/mobile-marketing-tools/latest-mobile-stats.

D. Doran (✉)
Department of Computer Science and Engineering, Kno.e.sis Research Center,
Wright State University, Dayton, OH 45435, USA
e-mail: derek.doran@wright.edu

V. Mendiratta
Bell Labs, Alcatel-Lucent, Naperville, IL 60563, USA
e-mail: veena.mendiratta@alcatel-lucent.com

D. Król et al. (eds.), *Propagation Phenomena in Real World Networks*,
Intelligent Systems Reference Library 85, DOI 10.1007/978-3-319-15916-4_11

led to the development of algorithms that automatically identify physical locations people are interested in [36] and reveal the typical mobility patterns of people within a country [5, 7, 41]. Studying the structure of calls placed between mobile devices have identified strong correlations between physical location and social friendship strength [14], and have even been used to discover regional economies within developing countries [34]. Such studies highlight the amazing ways mobile phone datasets let us study the collective actions of people through the structure of people's communications, interactions, and friendships. We have only just started to tap into the intelligence that can be mined from these datasets.

The main function of a mobile phone is to transfer information from one user to another. This information may be contained in the informal and unstructured data users transmit via SMS messages and voice calls. It may also be formal, structured data like images, files, and video transferred between devices in local areas through near-field communication (NFC) and bluetooth radios, or across the Internet to our contacts through smartphone apps and other third party services. Records about these transmissions are typically stored on a mobile device and may be collected by smartphone applications running in the background, or recorded by the network service provider. These records may reveal who information was transmitted to, what type of data was transferred, where the sender physically performed the transmission, and when the data transfer occurred. The relational nature of this data naturally gives rise to *networks* of users or devices within which many kinds of information flow. Since mobile phones are now ubiquitous across the world, understanding the process through which information propagates [29] across these networks adds to our basic understanding of the modern communication patterns humans exhibit.

In this chapter, we present a number of state-of-the-art propagation models and algorithms that have been applied to networks extracted from mobile phone datasets. The methods were selected so as to demonstrate the diversity of models that have been developed for this purpose, and to highlight the way they support many different innovative applications. We first discuss models that support the study of information diffusion across society. We then present epidemiological models that are tailored to the unique dynamics of communication between mobile devices in local areas, and how they are applied to anticipate the dynamics of malware transference between devices in local-area networks. Finally, we introduce sender-specific, receiver-specific, and clustering algorithms that compute the spread of information or influence and support a host of network provider services, including the identification of scammers and to predict who is likely to switch providers in the near future. We emphasize that this chapter is not meant to be a comprehensive survey of mobile phone data analytics, nor is it meant to present an exhaustive summary of the many propagation models that have been developed and could be utilized to understand mobile phone datasets. Instead, it intends to: (i) demonstrate how modeling propagation phenomena is a critical tool for mobile phone data analytics; (ii) show researchers interested in mobile phone data analytics the kinds of propagation models and algorithms they should be equipped with; and (iii) expose a number of avenues of future research in the study of propagation within mobile phone datasets.

This chapter is organized as follows. Section 11.2 introduces the kind of data and networks that may be extracted from mobile phone communications. Section 11.3 presents propagation models used to understand the diffusion of information across mobile phone networks. Section 11.4 discusses epidemiological models and their application to the study of mobile malware. Section 11.5 introduces propagation models used in the development of novel applications for service providers. Section 11.6 reflects on the works presented and offers exciting directions for future research. Concluding remarks are given in Sect. 11.7.

11.2 Mobile Phone Data Analytics

We define mobile phone data analytics as the mining and analysis of datasets whose records encode communication or interaction activities between mobile phone devices. Such datasets are typically extracted from a collection of devices that individually contain information about who the device's owner (i.e. mobile phone *user*) has a relationship with, as defined by the collection of mobile phone numbers in its contact list. The devices may also carry information about when and to whom the user transmits information via NFC or bluetooth to neighboring devices, and records of the SMS messages and phone calls she placed.

Smartphone applications that have sufficient permissions to access a device's data may extract information for performing mobile phone data analytics. Because it is difficult to deploy and obtain permissions for retrieving such information, however, researchers typically rely on call data records (CDRs) provided by a mobile phone service provider. The kind of information encoded in a typical CDR is provided in Table 11.1. It includes the phone number of the caller and callee, the duration of the call, the cost of making the call, if the call was on or off the provider's network, the date and duration of the call, and the base station used to connect the caller's mobile phone to the network. The position of this base station is used in many studies to approximate the position of a user when they make a phone call, while the duration, cost, and whether the call was on network may be attributes reflecting the strength of a relationship between two individuals. For example, we may infer that the back and forth off-network calls recorded in entries 1 and 2 of Table 11.1 represent communication between users who share a strong relationship since they both incurred a financial cost and spoke for a long period of time. The `calling_num` and `called_num` fields may be used to create a directed network of mobile phone calls between users.

The data collected from a mobile device or by a service provider may capture the structure of communications and relationships at multiple levels as illustrated in Fig. 11.1. At the *local level*, mobile devices equipped with NFC or bluetooth technology are capable to transmitting data between each other. At this level, the analysis exploits the position of devices to define a structure of possible local data transmissions to discover how data propagates in a small public area. These data transmissions may correspond to the automatic pinging of neighboring bluetooth

Table 11.1 Typical format and entries of a CDR

Call	Base_station	Calling_num	Called_num	Start_time	Duration	Cost	On_net_call
1	nyc-1234	8881112234	9992223345	01/01/2014 14:35:23	38	3.80	FALSE
2	paris-2512	9992223345	8881112234	01/03/2014 18:35:23	100	10.00	FALSE
3	chicago-3412	8882345678	8883345722	01/03/2014 18:40:30	50	0.00	TRUE

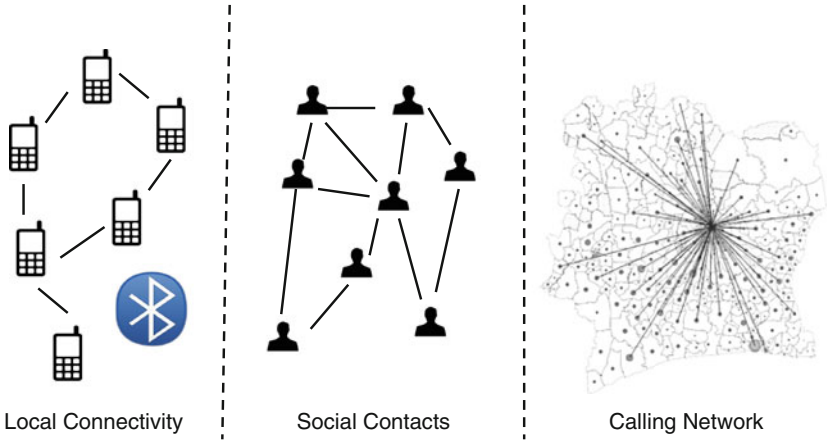

Fig. 11.1 Structure within mobile phone datasets among devices (local level), address books (contact level), and network-wide communication (calling level)

devices for deriving the density of people in an environment or to infer real-life social networks [37], data transmissions by an intentionally installed application, or the automatic spreading of malware or viruses that run without the user knowing [51]. At the *contact level*, phone numbers collected from the address book of users' devices are extracted and aggregated to form a collection of social relationships among users. At the *calling level*, CDRs collected by service providers may be used to study human communication across large geographic areas.

Many different propagation models and algorithms are applied to mobile phone data at the local, contact, and calling level. We divide the models covered in this chapter according to the type of analytics they support in Fig. 11.2, namely by: (i) understanding information diffusion; (ii) modeling malware propagation; and (iii) supporting network provider applications. These three types represent the diversity of the different kinds of mobile phone analytics supported by propagation models. They range from academic studies that seek to discover intrinsic qualities about information dissemination, to theoretical analyses that can be used to solve a widely-applicable problem facing society, to models that are specifically developed to support a business enterprise.

A roadmap of the specific models presented in this chapter is listed in Table 11.2, including a brief summary of the model and the network level it operates on. Information diffusion studies rely on structural models that capture spreading dynamics (causality trees), statistical approaches for characterizing complex distributions (mixture models and correlation metrics), and algorithms for finding users who play a critical role in the diffusion process (user clustering). The analysis of malware uses carefully designed SIR, SIS, and SIDR epidemiological models that also incorporate the unique mobility dynamics of mobile phone devices in public spaces. Practical applications utilize user clustering algorithms and new models for energy propagation across a mobile phone network.

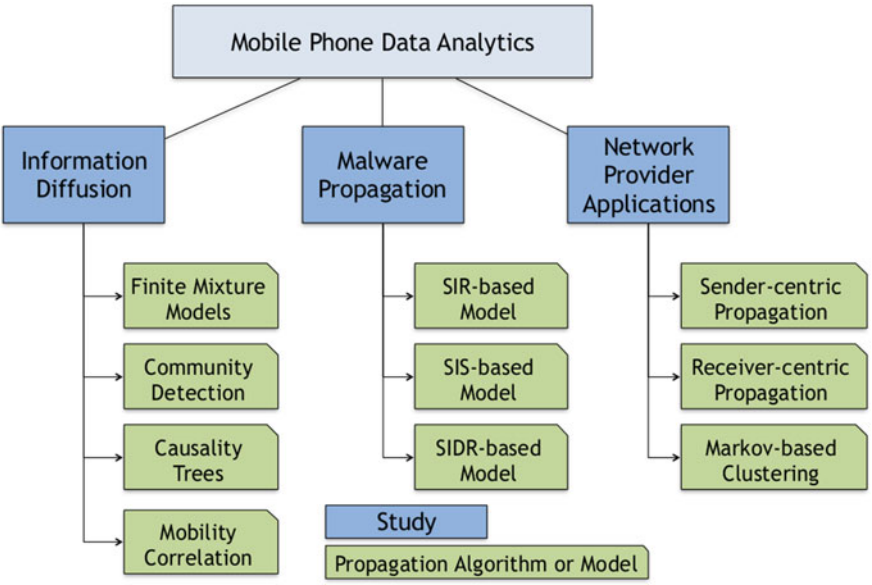

Fig. 11.2 Roadmap of the propagation models and the studies they support in this chapter

11.3 Information Diffusion

No matter the medium used to transmit information between mobile phone devices and their users, the chance that information spreads from one user to another depends on the strength of the relationship they share and on the dynamic nature of the information as it passes through the network of mobile users. Intuitively, the strength of the relationship shared between two users strongly impacts when, how often, and what kind of information is shared. Calls to a family member, for example, may happen much more frequently compared to calls made to a bank or doctor's office, increasing the chance of meaningful information dissemination. The dynamic nature of different types of information as it passes through a network is also critical. For example, information about topical news stories may experience a large number of transmissions due to the 'buzz' surrounding breaking news, but the frequency of these transmissions may decay over time as this news becomes less relevant. As another example, a person may broadly share a major life event with all of their contacts, but share a more personal story to a small subset of her contacts. We next examine information propagation models and algorithms that incorporates either of these aspects to make discoveries about the nature of mobile phone communication patterns.

Table 11.2 Propagation models and methods in this chapter

Propagation model	Summary	Structure level
Information diffusion (Sect. 11.3)		
Finite mixture model	Model the distribution of times between and duration of diffusions between links	Network
Mobility correlation metrics	Search for relationships between physical mobility and creating new network connections	Network
Causality trees	Models *pass-along dynamics* where information transmissions can only happen within a window of time τ	Contact
Community based greedy algorithm	Identify most influential members of a phone network under a weighted influence diffusion model	Network
Malware propagation (Sect. 11.4)		
SIR-based model	Models malware spreading dynamics where devices can recover and immunize themselves from infection	Local
SIS-based model	Model steady-state infection levels of malware in local areas where devices cannot be immune from infection	Local
SIDR-based model	Optimize the maximum damage that may be caused by a malware epidemic that not only infects but also kills devices	Local
Novel applications (Sect. 11.5)		
Sender-centric energy propagation	Model accumulation of influence where senders force information on receivers	Network
Receiver-centric energy propagation	Model accumulation of influence where receivers deicide what information is retained	Network
Markov clustering algorithm	Discover fraudulent users based on the structure of information propagation	Network

11.3.1 Characterizing Diffusion Frequency: Finite Mixture Models

One of the most basic properties of communication patterns are the frequency with which transmissions are made between users. Kim et al. [26] performed a comprehensive analysis of these frequencies by analyzing the communication activity of over one million bi-directional pairs of mobile phone subscribers from a nationwide cellular provider. Using metadata about each subscriber, they classified pairs by whether they are both in-network, if they are in different networks (out-network), and if they are family members. The objective of their study is to develop a universal model that can accurately capture the frequency of information exchange across all

three classes of users, as characterized by the inter-arrival times between calls made between pairs.

An initial analysis by the authors revealed that the empirical distribution of inter-arrival times do not follow a single exponential distribution, suggesting that the call arrival process is not Poisson for at least one class of pairs. They thus propose a finite mixture model to universally characterize the inter-arrival times of all pairs. A mixture model assumes that the data is drawn from a finite number of K distributions as specified by:

$$f(\mathbf{y};\psi) = \prod_{i=1}^{n} f(y_i,\psi) = \prod_{i=1}^{n}\sum_{k=1}^{K} w_k f_k(y_i;\theta_k) \tag{11.1}$$

where $f_k(y_i;\theta_k)$ is one of the K distributions of the mixture, $\mathbf{y} = (y_1,\ldots,y_n)$ is the vector of observations, and $w_1,\ldots,w_k$ are positive mixing weights assigned such that $\sum_{k=1}^{K} w_k = 1$. They decide to consider mixture models of Gamma, Lognormal, and Gaussian distributions because they all are capable of modeling non-negative random variables with a large range of possible density shapes. The model's collection of parameters ψ can be estimated by the expectation-maximization algorithm and use the Akaike Information Criterion [27] and Minimum Description Length [4] metrics to find the best number of components K.

11.3.1.1 Model Application

The authors fitted Gamma, Lognormal, and Gaussian finite mixture models to the distribution of inter-arrival times across all pairs of users and within the three different types of pairs. Although each pair of users exhibit a unique calling pattern, they find that the lognormal mixture model offers a very tight fit (MSE = 0.3605×10^{-4}). Family pairs were found to require a mixture model that is of higher order for fitting their inter-arrival time distributions, but of lower order to fit their call duration distribution. In contrast, out-of-network pairs need a low order mixture model to capture inter-arrival times and high order model to capture call durations. About 27 % of all pairs' inter-arrival time distributions are best fitted by a single order model.

11.3.2 User Mobility and Diffusion: Mobility Correlation Metrics

The distribution of peoples' physical locations are intimately related to the way information diffuses among users of a mobile phone network. This is because, practically, information passed through mediums like NFC or bluetooth require devices to be near each other. Furthermore, sociological studies confirm how we are more likely to connect and share information with those near us because the social links

encouraging this behavior are driven by spatial proximity [43]. Understanding the way humans diffuse physically is thus an important consideration when studying the spread of information across a mobile phone network.

Call data records record the id of the cell phone tower used by a sender and receiver used during a conversation. By mapping these id's to the physical position of the tower, we can study the approximate locations where a user regularly submits mobile phone calls and their daily trajectories through a geographic area. We can also find correlations between the physical proximity of two users and frequency of calls made between them. Such correlations can be expressed using a variety of metrics proposed by Wang et al. [48]:

1. *Distance*. This metric refers to the most likely physical distance separating two users in the network. Let $L_i(x)$ be the location of user x during his ith recorded call and $n(x)$ be the total number of calls made by x. Let

$$PV(x, l) = \sum_{i=1}^{n(x)} \mathbb{1}(l = L_i(x))/n(x) \tag{11.2}$$

be the probability that a user x visits a location l where $\mathbb{1}(q)$ is an indicator function that returns 1 if the statement q evaluates to true and 0 otherwise. The *most likely* location of user x is thus given by $ML(x) = \arg\max_{l \in Loc} PV(x, l)$. We can define the distance d between users x and y as $d(x, y) = \text{dist}(ML(x), ML(y))$ where dist is a measure of geographic distance.
2. *Spatial Co-location rate*. This metric captures the likelihood that two users visit in the same location but not necessarily at the same time. Assuming their visits are independent, it is given as:

$$CoL(x, y) = \sum_{l \in Loc} PV(x, l) \times PV(y, l) \tag{11.3}$$

where Loc is the set of locations that both x and y have been recorded as visiting.
3. *Cosine similarity*. This metric uses cosine similarity to capture how similarly two users frequent the same locations. It is given as:

$$Cos(x, y) = \sum_{l \in Loc} \frac{CoL(x, y)}{||PV(x, l)|| \times ||PV(y, l)||} \tag{11.4}$$

4. *Weighted cosine similarity*. This metric corresponds to the *tf-idf* version of cosine similarity. In essence, the *tf-idf* version adds weight to co-location events within low-density areas, that is, areas where users are seldom seen, and penalizes high-density areas. For example, pairs that frequent seldom visited locations may be more likely to have a relation than those who both frequent common locations.
5. *Co-location rate*. This metric measures the probability two users will be located in the same location in the same day and hour. It is given as:

$$CoL = \frac{\sum_{i=1}^{n(x)} \sum_{j=1}^{n(y)} \theta(\Delta T - |T_i(x) - T_j(y)|) \mathbb{1}(L_i(x) = L_j(y))}{\sum_{i=1}^{n(x)} \sum_{j=1}^{n(y)} \theta(\Delta T - |T_i(x) - T_j(y)|)} \tag{11.5}$$

where $\theta(x)$ is the Heaviside step function and $\Delta T = 1$ h. The numerator counts the number of times two users visit the same location at the same time, normalized by how frequently they are active at the same time.

6. *Weighted Co-location rate*. This is the *tf-idf* version of *CoL* where the normalization factor is the log of the number of users at each location in the same hour.
7. *Extra-role Co-location rate*. This metric is defined by *CoL* taken over only evening and weekend hours. Co-location during these times may be an important predictor of an offline relationship.

11.3.2.1 Model Application

Wang et al. applied these mobility correlation metrics to a dataset consisting of over 6 million users and 90 million calls [48]. Their analysis focuses on the 50,000 most active individuals in the dataset. They find that the geographical distances between pairs exhibit a heavy-tailed distribution, which is consistent with a number of previous findings [28, 31, 33]. The CoL and SCoL measures of co-location rates reveal how many pairs can be found to be visiting the same locations, but for short periods of time. Furthermore, the geographical distance between two users decays only logarithmically with the *Col* and *Cos* measures of proximity.

Since mobility and information diffusion are intimately related to each other, the authors utilize these metrics to predict whether new information diffusions will occur in the future. They train a C4.5 decision tree to classify whether a potential connection in the calling network that does not exist during time period t will emerge at time period $t + 1$. The tree is trained with network structure and mobility correlation metrics and yields a precision of 73.5 % and recall of 66.1 %. Compared to classifiers that only consider network structure metrics, this precision and recall is an order of magnitude higher. This confirms that human mobility patterns are intimately associated with the future diffusion of information across new connections.

11.3.3 Modeling Pass-Along Dynamics: Causality Trees

An intriguing type of information people share between both their peers and close contacts are breaking news stories or rumors. We define such information to be *short-lived*, as people become disinterested in news and rumors the longer it has been since it broke out on the network. To model the dissemination of such information, we consider *pass-along spreading processes* [39]. A pass-along spreading process is defined as one where a user can only pass information to some subset of their contacts, and only within a short a period of time τ since she received the information. This

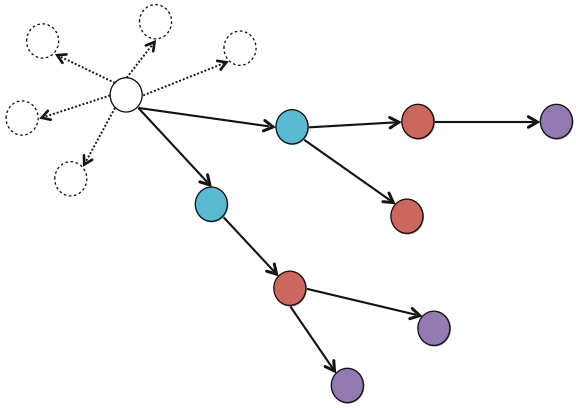

Fig. 11.3 Example of a pass-along dynamic modeled by a causality tree where $d = 3$ and $s = 9$. The root user of the tree passes information along to $k_o = 2$ out of his $k'_o = 7$ contacts. Users with the same color exist at the same depth of the tree

pass-along process repeats for every user that has received this information, until no new users have become informed. Figure 11.3 illustrates how a pass-along process is modeled as a diffusion tree whose depth d corresponds to the maximum distance from the initiator to an informed user, size s is the number of users who become informed, and whose paths represent a sequence of consecutive communications whose time between calls are always less than or equal to τ.

A causality tree can be used to model the probability a user k will be contacted by k_i other users and subsequently pass along information to k_o users within a given τ. Such an event corresponds to a user in a causality tree that has in-degree k_i and out-degree k_o given τ. These probabilities can be used to identify the extent to which a user in the network chooses to participate in the pass-along process. For example, a user who is entirely disinterested in spreading information would be represented in the model as a user in the tree with large in-degree and low out-degree. Users excited to pass information widely corresponds to those having large out-degrees in the cascade tree. Let k'_i and k'_o be the in- and out-degree of node k across a network of contacts (e.g., the number of others who have k as a contact and number of contacts k has, respectively). Since k receives and sends information from and to only a subset of all contacts during a pass-along along process, the probability k has in-degree k_i and out-degree k_o in a causality tree is given by:

$$p(k_i, k_o; \tau) = \sum_{k'_i=k_i; k'_o=k_o}^{\infty} p_\infty(k'_i, k'_o) \times \binom{k'_i}{k_i} T_i(k'_i, \tau)^k (1 - T_i(k'_i, \tau))^{k'_i - k_i} \times \binom{k'_o}{k_o} T_o(k'_o, \tau)^k (1 - T_o(k'_o, \tau))^{k'_o - k_o} \tag{11.6}$$

where $p_\infty(i, o)$ is the probability of finding a node with in-degree i and out-degree o in the contact network and $T_o(k_i', \tau)$ ($T_i(k_i', \tau)$) is the probability that a user will send (receive) information to (from) k_i' (k_o') users within τ time. We can simplify Eq. 11.6 by assuming that the number of users k chooses to send information to is independent of the number of sources k received the information from, so that $T_i(k_i', \tau) = T_o(k_o', \tau) = T(k, \tau)$. If we assume that the frequency with which calls are made over a communication link follow a Poisson process [47], we can model the probability that k will send short-lived information to a contact within τ time as $1 - \exp(-\rho\,\tau)$, where ρ is defined as the sending rate of k. Thus, we can define $T(k, \tau)$ as:

$$T(k, \tau) = \int d\rho\, p(\rho)(1 - \exp(-\rho\tau)) \tag{11.7}$$

where $p(\rho)$ is the probability density of user sending rates across the network.

While $p(k_i, k_o; \tau)$ represents the dynamics of individuals in a pass-along process, statistics about the causality tree itself sheds light into the overall reach and participation of users sharing short-lived information. The recursive nature of a cascade tree can be exploited for this purpose. For example, to compute the probability of observing a tree with size s $p(s; \tau)$, we begin by defining the probability of finding a tree of size $s = 1$ by $p(s = 1; \tau) = p(k_o = 0; \tau)$, i.e., the probability of a tree whose root node has out-degree zero. $p(s = 2; \tau)$ can then be defined as the probability that a root node has out-degree 1 and its child node also has out-degree 1. We can continue to extend this definition recursively to define all $p(s'; \tau)$ for $s' < s$. This recursive relationship may be expressed by the generating function $G(z, \tau) = E(z^s) = \sum_{s=1} p(s; \tau) z^s$, which obeys the self-consistency equation:

$$G(z; \tau) = zg(1, G(z; \tau); \tau) \tag{11.8}$$

where $g(1, y; \tau)$ is the generating function for the probability a user in a cascade tree has out-degree k_i:

$$g(1, y; \tau) = \sum_{k_o} p(k_o; \tau) y^{k_o} \tag{11.9}$$

The cascade size distribution can thus be found by taking derivatives of the generating function:

$$p(s = n; \tau) = \frac{1}{n!} \frac{\partial^n G(z; \tau)}{\partial z^n}|_{z=0} \tag{11.10}$$

A similar recursive formulation can be used to model the probability a tree has depth d $p(d; \tau)$. Let $E_d(\tau)$ be the probability a causality has some depth less than or equal to d. This probability obeys the relation:

$$E_d(\tau) = g_1(E_{d-1}(\tau); \tau) = g_d(0; \tau) \tag{11.11}$$

where $g_1(y; \tau) = g(1, y\,\tau)$ and $g_n(y; \tau) = g_1(g_{n-1}(y; \tau); \tau)$. Then the probability of a tree having depth d is given as:

$$p(d; \tau) = E_d(\tau) - E_{d-1}(\tau) = g_d(0; \tau) - g_{d-1}(0; \tau) \tag{11.12}$$

11.3.3.1 Model Application

Peruani et al. proposed the propagation model based on causality tree presented above [39]. They applied it to a mobile phone dataset from a European telecom with 1,044,397 users that made 13,983,433 calls between them. They derive the parameters of the model from the dataset, and identify a very close fit between the modeled cascade size and probability distributions with the observations they make in the original dataset.

The model's application draws a number of findings about the nature of pass-along dynamics in a mobile phone network. Specifically, they find the existence of super-spreaders and receivers, who are giant hubs that absorb or widely disseminate information along the network. They also discover that pass-along dynamics are extremely sensitive to the correlation of users' in- and out-degree distributions. Furthermore, at large time-scales (τ), the spreading dynamics actually become dominated by correlations in the topological structure of users in the network, not the pass-along process. In other words, pass-along processes only capture the dynamics of information exchange at a very local level (e.g. to degree 1 or 2-neighbors).

11.3.4 Diffusion Maximization: Community Based Greedy Algorithm

A third-party wishing to influence as many people as possible may wish to find k seed nodes who can maximize the spread of their influential information across the network. These seed nodes represent *influential users*, defined as those who share information with the intention of changing another's personal opinions or beliefs. If an influencer is successful, newly influenced people subsequently pass their information off through their set of connections, and so forth. Influence propagation thus exhibits the same pass-along dynamics modeled by causality trees, but without a time constraint. In other words, an influencer may try to sway another at any time, regardless of the time passed since they themselves became influenced. The extent to which influence propagates through a network thus depends only on the position and number of influencers that begin the diffusion process.

Although finding the k users to initially influence such that the maximum number of others on the network become influenced is an NP-hard problem, greedy algorithms are capable of finding an approximate solution to within a factor of $(1 - 1/e - \varepsilon)$ [8, 21, 32], the algorithms are too inefficient to process very large mobile phone networks. Instead, community-based greedy algorithms that identify the top-k most influential nodes in a mobile phone network have been proposed as a way to efficiently solve this problem [49]. We first define the *diffusion speed* of information from user v_i to v_j in the network as:

$$\lambda_{ij} = 2\bar{\lambda}\frac{w_{ij}}{w_{max} + w_{min}} \tag{11.13}$$

where $\bar{\lambda}$ is the empirically measured average calling rate of users in a network and w_{ij} is the weight of the directed connection from i to j. These weights should correspond to a quality of the connection such that the higher its value, the faster the rate of information diffusion. For example, the number of calls or SMS messages sent between the users could correspond to a connection weight. The algorithm then considers the following diffusion process:

1. Select a set of active seed nodes S_0 active at an initial time $t = 0$.
2. Increment the time clock to $t = t+1$. Choose a node v_i from the set S_{t-1}. For every directed neighbor v_j of v_i, try to influence her with probability λ_{ij}. If successful, add v_j to the set S_t.
3. Update $S_t = S_t \cup (S_{t-1} \setminus v_i)$.
4. Repeat steps 2 and 3 until the set of active nodes $S_t = \emptyset$.
5. The set of all nodes influenced by the seed set S_0 is given as $\mathcal{V}_S = \bigcup_{i=0}^{t-1} S_i$. Define the *degree of influence* of S_0 to be $R(S_0) = \mathcal{V}_S/N$ where N is the number of users in the mobile phone network.

Under this process, we can efficiently find a set of seed nodes S such that $|S| = k$ if we assume that the mobile phone network can be divided into many communities of users. A community is a set of users who frequently communicate with each other and are more likely to be swayed by information originating within it. If information originating from one community will have almost no influence over the members of another, a good approximation for finding the top influencers of the network is to find users *within individual communities* that maximize the spread of influence within them. The algorithm for finding these communities are given in [49].

Let I_k be the set of the k seed users that leaves the strongest amount of influence on the network. To find I_k, assume that we already have constructed the set I_{k-1} thus far. We define by how much the degree of influence across the network will increase by adding the most influential member within community C_m to the set I_k as:

$$\Delta R_m = \max\{R_m(I_{k-1} \cup v_j) - R_m(I_{k-1}) | v_j \in C_m\} \tag{11.14}$$

Thus, we can choose the kth influential user to add to I_{k-1} by choosing the most influential member in the community that has the largest ΔR_m value. ΔR_m can be found using any previously proposed less efficient algorithm to find the most influential user within a small network [8]. This less efficient algorithm is expected to perform in a reasonable amount of time since it only runs across a community of the entire calling network.

We can use dynamic programming to efficiently choose the community from which an influential user is added I_k. Let $R[m, k]$ be the influence degree yielded if the kth most influential user is selected from one of the first m communities. Then,

$$R[m, k] = \max(R[m-1, k], R[m, k-1] + \Delta R_m) \tag{11.15}$$

where $R[m, 0] = 0$ and $R[0, k] = 0$. In other words, if a user from the first $m-1$ communities yields a smaller influence degree than choosing the most influential user from community m, choose it from C_m. Otherwise, choose it from one of the $m-1$ former communities. The choice of these former communities is represented by $s[m, k]$. It is given by:

$$s[m, k] = \begin{cases} s[m-1, k], & R[m-1, k] \geq R[M, k-1] + \Delta R_m \\ m, & R[m-1, k] < R[m, k-1] + \Delta R_m \end{cases} \tag{11.16}$$

with $s[0, k] = 0$.

11.3.4.1 Model Application

Wang et al. presented and applied this community-based greedy algorithm to a network of SMS messages between 723,201 users collected by a major telecom company [49]. Under many choices of K and $\bar{\lambda}$, the community-based method was able to find a set of users that yields the largest spread of influence in the network compared to many previously proposed algorithms. It has modest run-times (on the order of thousands of seconds) under the entire range of parameter settings used for experimental analysis under a simple hardware configuration (2.0 GHz Xeon 8 Core CPU; 8GB Memory; Debian 4.0 Operating System). Experimental analysis finds that the improvement in influence degree rises exponentially fast with $\bar{\lambda}$ (the average rate of diffusion). Influence degree increases just logarithmically with K, with very small gains for $K > 15$. The study also finds that approximately $M = 25$ communities offers the best tradeoff between minimizing computation time and maximizing influence degree. In summary, the method demonstrates how a small number of influencers ($\sim$15) are sufficient to widely disseminate influence across society-wide communication networks. Furthermore, numerous latent communities exist within a mobile phone network, where members are likely to influence each other.

11.4 Malware Propagation

Security and network researchers envision mobile phones as being the next frontier for malware [9, 10, 15] due to the many vulnerabilities present in mobile platforms [20], the un-savvy users operating mobile devices, and the private and valuable information they store on them. A 2011 Mobile Threats Report by Juniper Networks Mobile Threat Center found a 155 % increase in mobile malware over the past year [45]; by the end of the same year McAfee Labs had collected over 75 million samples of mobile malware. Malware is capable of changing mobile phone configurations, spamming SMS messages, dialing pay-to-call numbers, and collecting private information stored on the device.

Understanding the development of malware on a mobile phone network, and devising techniques to combat this threat, require novel propagation models. This is because these always-on devices may be susceptible to infection through local NFC or bluetooth transmissions , by connecting to a compromised public access point, or through a compromised link shared across a contact network via SMS [38]. These infections may thus quickly propagate through a mobile network as it infects and transmits from device to device. In many ways, this is analogous to the spread of an infectious disease through a population of people who congregate in public places. Thus, many researchers have proposed different variations of common epidemiological models (e.g. SI [3], SIR [23], SIS [24]) to better understand the spreading dynamics of mobile phone malware, and to propose methods that thwart their spread. This section details some of these recent models and methods, and discusses their application to mobile phone networks.

11.4.1 Infection Dynamics with Recoverable Devices: SIR Epidemiological Model

Rhodes et al. introduced an extended SIR epidemiological model for modeling the spread of malware opportunistically shared between bluetooth enabled smartphones [42]. The model considers not just the rate at which devices become **susceptible** (**S**), **infected** (**I**), or **recovered** (**R**), but also the rate at which devices come into contact with each other and the devices' transmission profiles. We first assume that mobile devices are spatially distributed over a fixed region with density ρ. Each individual device moves independently of all others with constant velocity v. If any device moves within the transmission radius R of another device in the area, the devices make contact and there is an opportunity for malware to spread. Thus, a new individual device that moves with its own velocity v_i will be exposed to contact by device i during a time period dt if it lies within a rectangular-shaped area that is covered by the movement of i and lies in the direction of the vector $w = v_i - v$. The total area covered by i during dt is given by $dA = 2Rwdt$ where $w = \sqrt{v_i^2 + v^2 - 2v_i \cos\phi}$ is the relative speed of the device and ϕ is the angle between velocity vectors. Thus, the number of devices in transmission range of i is given by:

$$\gamma = \int_0^{2\pi} \frac{dN_\phi}{dt} = \frac{\rho R}{\pi} \int_0^{2\pi} w d\,\phi \tag{11.17}$$

This reduces to:

$$\gamma = \frac{4\,\rho R}{\pi}\,(v_i + v) \int_0^{\pi/2} \left(1 - \frac{4vv_i}{(v + v_i)^2} \sin^2 \omega\right)^{1/2} d\omega \tag{11.18}$$

If we make the simplifying assumption that the new device i moves with the same velocity as all other devices (so that $v_i = v$), we can write Eq. 11.18 as an elliptic integral and use its standard form to find:

$$\gamma = \frac{8}{\pi}\,\rho\, vR \tag{11.19}$$

If a single device transmits malware to another within its range with probability p, the infection rate of devices in the system is $\beta = p\,\gamma$.

The model also considers a radial decay function to compute the probability a susceptible device becomes infected. The choice of a radial decay function is based on the fact that the longer a device spends in the transmission range of an infected user, the higher its chance of becoming infected, and the closer one device is to another, the longer it will take for them to be out of transmission range. Thus, we compute the probability a device at position r gets transmitted malware by computing the path length between r and contact with an infected node given by $2(R^2 - r^2)^{1/2}$, multiply it by the probability of infection upon falling in transmission range p, and normalize by the total transmission range:

$$p(r) = \frac{p}{R}(R^2 - r^2)^{1/2} \tag{11.20}$$

Integrating over all positions r and substituting p and R for $p(r)$ in the formulation of β, we get:

$$\beta = \frac{8}{\pi}\,\rho\, v \int_0^R p(r) dr \tag{11.21}$$

which solves to:

$$\beta = 2R\,\rho\, vp \tag{11.22}$$

Using this new infection rate, we apply the SIR model to specify a malware outbreak by the differential equations:

$$\frac{d\mathbf{S}}{dt} = -\beta \frac{\mathbf{SI}}{N} \tag{11.23}$$

$$\frac{d\mathbf{I}}{dt} = \beta \frac{\mathbf{SI}}{N} - \delta \mathbf{I} \tag{11.24}$$

$$\frac{d\mathbf{R}}{dt} = \delta \mathbf{I} \tag{11.25}$$

where I is the number of devices infected, S is the number of susceptible devices, and N is the total number of devices on the network.

11.4.1.1 Model Application

The authors compared the output of the SIR-based model to a simulation of an outbreak of malware in a setting with a device density of 3000 devices/km^2, mean velocity of 2 km/day, transmission probability $p = 0.1$, transmission range of 5–40 m per device, and with a recovery rate of 1 device per 5 days. They find that the epidemic dynamics are mostly caused by the aggregation of many dyadic interactions, rather than spreading the malware to multiple devices at once due to the the small transmission range of the devices. However, as transmission radius increases, the SIS-model comes to a much stronger agreement with the simulation results. They conclude that the dynamics of malware propagation are greatly affected by the characteristics of the devices and of the environment they operate under. When malware that devices can recover from are transmitted over far-reaching channels, the SIS-model captures its infection dynamics very well.

11.4.2 Infection Dynamics Without Immunization: SIS Epidemiological Model

Mickens et al. developed an extension of the Kephart-White (KW) epidemiological model [22] that also considers the mobility of devices within a constrained area [35]. This is an SIS (Susceptible–Infected–Susceptible) epidemiological model where devices may cycle between **susceptible** and **infected**. In other words, a device can never be completely immune and may become infected again once cured.

The traditional KW model assumes a homogeneous network topology in which all devices have a similar number of neighbors $\bar{k}$. If I is the fraction of devices infected at a particular moment in time, the KW model describes the propagation of an infection as the differential equation:

$$\frac{dI}{dt} = \beta \bar{k} I(1 - I) - \delta I \tag{11.26}$$

where t is the current time, β is the propagation rate of malware from one device to another, and δ is the rate at which any infected device is cured. Holding these rates and $\bar{k}$ constant, this equation has a steady state solution of:

$$I = 1 - \frac{\delta}{\beta \bar{k}} \tag{11.27}$$

Thus, we require

$$\beta \bar{k} > \delta \tag{11.28}$$

for an infection to persist in the network. These parameters can be mapped to model the spread of mobile device malware by letting $\bar{k}$ be the average number of devices within communication range of any other device, β be the probability a malware infected device transmits it to a health neighbor during a time period Δt, and δ be the probability an infected device cures itself during time Δt. However, extensive analysis by the authors confirm that the KW model does not accurately model the dynamics of malware that spreads by NFC or bluetooth transmissions in a local mobile phone network. This is because the homogeneity assumption held by the KW model is broken by the fact that mobile devices move around a region and have a limited transmission radius. The number of neighbors a device has at any given time is thus constantly in flux and should not be represented by a constant value $\bar{k}$. Furthermore, the KW model does not incorporate parameters for the velocity of mobile devices within an area, which they find to be a major factor in how quickly malware spreads in their simulations.

To extend the KW model, the authors consider the spatiotemporal dynamics of devices within a large rectangular area using a random waypoint mobility model. In this mobility model, devices randomly select a destination point, travel there, pause for a constant time t_p, and then choose another random destination point. The waypoints are independently chosen prior to departing. The speed at which devices move between waypoints is given as a random velocity chosen uniformly within some pre-specified range. Under this mobility model, the spatial density function of devices over a square region is given as:

$$S(x, y) = \frac{p_p}{a^2} + (1 - p_p)\frac{36}{a^6}(x^2 - \frac{a^2}{4})(y^2 - \frac{a^2}{4}) \tag{11.29}$$

where a is the length of a side of the square region and $p_p = t_p/E[T]$ where $E[T]$ is the average time a node takes to move from one waypoint to another. Thus, if a device is at position (x_i, y_i), we can derive the probability that it is within communication range of another device by the integral:

$$c(x_i, y_i) = \int_{y_i - r}^{y_i + r} \int_{x_i - \sqrt{r^2 - (y - y_i)^2}}^{x_i + \sqrt{r^2 - (y - y_i)^2}} S(x, y) dx dy \tag{11.30}$$

where r is the radius of communication for all devices. We use c to find the probability a device at (x_i, y_i) has k_i devices within communication range by:

$$Pr(x, y, k = k_i) = \binom{N-1}{k} c(x, y)^k (1 - c(x, y))^{N-k-1} \tag{11.31}$$

The expected probability two devices will be within communication range of each other is thus:

$$\bar{c} = \frac{\int_{-a/2}^{a/2} \int_{-a/2}^{a/2} c(x, y) dxdy}{a^2} \tag{11.32}$$

and the probability any device will have k_i devices in communication range across the entire region is:

$$Pr(k = k_i) = \frac{\int_{-a/2}^{a/2} \int_{-a/2}^{a/2} Pr(x, y, k = k_i) dxdy}{a^2} \tag{11.33}$$

To consider mobility under the KW model, the connectivity fluctuations induced by mobility need to be incorporated. We can do so by considering the average travel time of a device from one waypoint to another $E[T]$ as a queue or pipe that takes $E[T]$ time to traverse. If the probability a device at any location has k_i neighbors is $Pr(k = k_i)$, the amount of time it spends with k_i neighbors while moving from one location to another is given by $E[T] \times Pr(k = k_i)$. For example, $E[T] \times Pr(k = k_0)$ is the amount of time a device has no neighbors while it travels from one destination to another, and hence can be subjected to malware cures. Otherwise, for $E[T] \times Pr(k = (k_i > 0))$ time units, the device is subject to an infection pressure proportional to $\beta\, k_i$ and a cure pressure proportional to δ. The extended KW model thus requires

$$\sum_{k_i=0}^{N-1} \beta\, k_i Pr(k = k_i) E[T] > c\, \delta\, E[T] \tag{11.34}$$

for a malware outbreak in the network to exist, where c is a constant account for global factors affecting connectivity. Since Eq. 11.33 tells us the percentage of time a device has k_i other neighbors, the total number of devices with k_i neighbors across the local area is given by $N \times Pr(k = k_i)$ where N is the number of devices in the local network.

To help compute the steady-state infection level of the mobile network, let us assume that the stretches of time a node has k_i neighbors are large relative to the unit of time used to measure infection rates Δt. Consider a collection of N queues $\{Q_{k_i}\}$, each of which initially has $N \times Pr(k = k_i)$ devices in it. When a device enters Q_{k_i}, it spends $E[T] \times Pr(k = k_i)$ time in it before exiting. Each queue can be thought of as a separate KW process described by the rates of infection β and curing δ, where all devices in the queue have the same $\bar{k} = k_i$ neighbors. Treating all devices in the

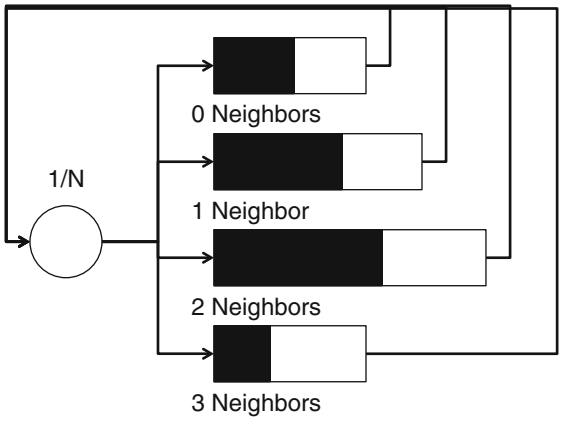

Fig. 11.4 Queueing network for finding steady-state infection levels. Each queue is loaded with devices that have the same number of neighbors, so queue i starts with $N \times Pr(k = k_i)$ devices. A random proportion of devices in queues (shown in *black*) are infected. At every time-step, we infect and cure devices according to a KW process that runs separately within each queue. After $E[T] \times Pr(k = k_i)$ time-steps, a device in queue i departs and is divided into $1/N$ units. These small units are then distributed across all of the queues

same queue under the same KW process is intuitive because they all have the same number of k_i neighbors, which is a core assumption of the KW model.

We can utilize a network of these queues, illustrated in Fig. 11.4, to find the steady-state infection levels. We initially place $N \times Pr(k = k_i)$ in each queue and assign a random proportion $I_{init} \in [0, 1]$ of its devices to be infected with malware. The model then iteratively updates itself in increments of Δt. At each update, it first simulates a propagation of the malware in each queue Q_{k_i} using the KW equation:

$$\frac{dI_{Q_{k_i}}}{dt} = \beta\, k_i I_{Q_{K_i}}(1 - I_{Q_{K_i}}) - \delta\, I_{Q_{K_i}} \tag{11.35}$$

Every Δt time units, the model checks if the exit time of any device has exceeded the current time, and if so, it removes the device from its queue, divides it into N equally sized pieces, and enqueue's one of these pieces into the rest of the queues. Finally, every queue updates its infected percentage $I_{Q_{k_i}}$ to reflect its newly enlarged population and infection percentages. At any moment during this process, the total number of infected devices in the network is given by:

$$\sum_{k_i=0}^{N-1} I_{Q_{k_i}} \times |Q_{k_i}| \tag{11.36}$$

where $|Q_{k_i}|$ is the number of devices in a queue. The steady state number of infected devices can be found by continuing to iterate the model until these values converge.

11.4.2.1 Model Application

Mickens et al. simulated a mobile device network where devices have a 100 m communication radius and move within a square region with 1000 m sides. Using various device velocities and number of devices in the network, they compare the predicted proportion of infections given under the KW model and their extended queueing based model against the simulation results. They find that the steady-state infections projected by the KW model was different from the simulation by 12.5 %, while the queueing model was only off by 4.0 %. Their analysis also discovers that epidemics are *unstable* under many parameter settings. For example, in five simulation runs lasting 200,000 s, one epidemic died out almost immediately, another lasted the entire time, and the others lasted between one- and three-fourths of the total simulation time. Thus, while the extended KW model accurately predicts average levels of infection, it hides the instability of the malware propagation process.

Finally, the authors apply their model to a scenario where the spatial distribution of devices across the region is strongly skewed, that is, where devices tend to favor specific areas within the region. This scenario may better reflect real-life mobility patterns, as users tend to congregate around popular landmarks within a region. The waypoint mobility model was modified so that nodes have a higher probability to travel to one of three 'hot-spots' in the square region. For different values of N, the queueing model outperforms the KW model in predicting the steady-state infection levels under the modified mobility model, but the relative improvement is not as large. The authors hypothesize that their queueing based model, which only captures the number of neighbors a device has at any time, does not necessarily capture the spatial distribution of devices within a geographic region.

11.4.3 Maximizing Malware Damage: SIDR Epidemiological Model

Khouzani et al. propose the analysis of an SIDR epidemiological model to estimate the maximum amount of damage malware can impart on a local mobile wireless network [25]. They define damage as a cumulative function that increases with the number of devices that may be *infected* or *dead*. Their model allows this damage function to be generally defined, and assumes that the malware wishes to maximize damage subject to specific constraints on the energy consumption of its host devices.

Under an SIDR model, devices may fall under one of four states: **susceptible** (**S**), where an unprotected device is not yet infected; **infective** (**I**), where a device has been loaded with malware, and may propagate it to others, but the malware has not yet attacked the device; **dead** (**D**), where the malware successfully compromised the device; and **recovered** (**R**), where an updated device is immune from the infection. We let $n_\alpha(t)$ be the number of devices in state $\alpha \in \{S, I, D, R\}$ such that $\sum_\alpha n_\alpha(t) = N$ is the number of devices in the model, and the proportion of all devices in each state

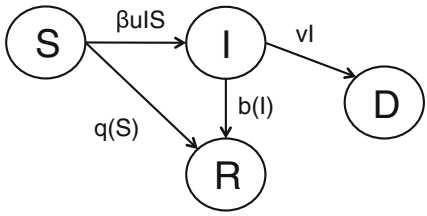

Fig. 11.5 Markov model and transition probabilities for device states under the SIDR model

as $S(t), I(t), D(t)$, and $R(t)$ respectively so that $S(t) + I(t) + D(t) + R(t) = 1$. We assume that an outbreak begins at time $t = 0$ with the infection of $I(0) = I_0$ devices. The initial conditions of the system are $R(0) = D(0) = 0$ and $S(0) = 1 - I(0)$.

Infections occur as devices within a region A move with velocity v. Infective devices transmit malware once they fall within a given transmission range. The probability of an infection is based on two factors: the *density* of devices within A, given as $\nu_1 = |N|/|A|$, and the rate at which a given pair of devices contact each other, given as $\nu_2 = 1/A$ [18]. If $u(t)$ is the product of an infected device's transmission range and rate at which it scans for devices to transmit to, the process of malware transmissions from an infected to susceptible device can be modeled by an exponential random process whose rate at time t is $\hat{\beta}u(t)$ where $\hat{\beta} = \nu_1\nu_2$. Infected devices will be killed after an exponentially distributed random amount of time with rate $v(t)$. An infected or susceptible device may also recover after infection by healing or immunizing itself with rates given by $B(I(t))$ and $Q(S(t))$, respectively. The rate functions B and Q can be defined in any way the modeler would like, as long as they meet the following criteria: (i) $\lim_{x\to 0} B(x) < \infty$ and $\lim_{x\to 0} Q(x) < \infty$; (ii) for $0 < x < 1$, B and Q are positive and differentiable; and (iii) $xB(x)$ is a concave non-decreasing function of x and $xq(x)$ is also a non-decreasing function of x.

Under these infection and recovery dynamics, we can model the rates at which devices transition between states using the continuous time Markov chain in Fig. 11.5. We represent the state vector of this chain as $V = (n_S(t), n_I(t), n_D(t))$, dropping $n_R(t)$ since $n_S(t) + n_I(t) + n_D(t) = 1 - n_R(t)$. Let $\beta = \lim_{N\to\infty} N\hat{\beta}$, $q(S) = Q(S)S$, and $b(I) = B(I)I$. According to [30], $S(t), I(t)$, and $D(t)$ will converge to the solution of the following differential equations as N grows:

$$\frac{dS(t)}{dt} = -\beta\, u(t)I(t)S(t) - q(S(t)) \qquad S(0) = 1 - I_0 \qquad (11.37)$$

$$\frac{dI(t)}{dt} = \beta\, uI(t)S(t) - b(I(t)) - v(t)I(t) \qquad I(0) = I_0 \qquad (11.38)$$

$$\frac{dD(t)}{dt} = v(t)I(t) \qquad D(0) = 0 \qquad (11.39)$$

These equations satisfy $0 \le S(t), I(t), D(t)$ and $S(t) + I(t) + D(t) \le 1$ for all t.

We now consider an attacker who wants to infect a local area in such a way that the amount of damage caused by the malware infection during a window of time

[0, T] is maximized. Since damage corresponds to both the infection and killing of devices in the network, the damage function can take the following general form:

$$J = \kappa D(T) + \int_0^T f(I(t))dt \tag{11.40}$$

κ is a positive 'reward' per device killed and f is an increasing convex function where $f(0) = 0$. An attacker will try to maximize J by regulating two parameters of the malware: the rate at which it will kill devices $v(t)$ and the product of the malware's transmission range and scanning rates $u(t)$. The choice of parameters for these values are subject to:

$$0 \leq v(t) \leq v_{max} \tag{11.41}$$

$$0 \leq u_{min} \leq u(t) \leq u_{max} \tag{11.42}$$

$$\int_0^T h(u(t))dt \leq C \tag{11.43}$$

The upper bound on $v(t)$ represents an inherit maximum speed at which a device can be killed by an infection. The bounds on $u(t)$ represent maximum transmission rates caused by the physical properties of an environment. The integral constraint over $h(u(t))$ ensures that the malware infection does not fully deplete an infected device's power, which it relies on to spread the infection and to eventually kill the device. It is assumed that h is a non-decreasing and non-negative function. Once the malware chooses v and u, the Markov chain's state vector V will be specified at all times t, allowing us to solve the system of differential equations and hence compute the damage J of the attack. We can then find optimal functions that control the killing, transmission, and scanning rates $v(t)$ and $u(t)$ to maximize J.

11.4.3.1 Model Application

Khouzani et al. studied the proposed SIDR model and damage function under various parameter settings to gain insights about the malware infection and recovery processes [25]. From the optimal forms of $v(t)$ and $u(t)$, they discover how malware should start with a small killing rate that gradually increases over time. This way, infected devices are given an opportunity to infect others before being killed off. When the time window is almost over, however, devices should adopt a high killing rate to take as many down as possible. Furthermore, the malware should not decrease an infected devices transmission and scanning rates until approximately one third of the time window has elapsed. If the network can increase the recovery $B(t)$ and immunization rates ($Q(t)$ and $B(t)$) of devices, the malware must extend the period during which its transmission and scanning rates are highest ($u(t) = u_{max}$).

The authors also find a relationship between recovery rates and the total damage imposed by malware. Interestingly, they find that the amount by which damage is reduced decreases exponentially with the rate of recovery. However, with larger recover rates comes larger bandwidth and power costs for the devices.

11.5 Novel Applications

Mobile service providers collect a wealth of information about their customers and their calling behaviors. Hidden within these records are patterns that may be exploited to help the provider offer better service to their customers, or to make discoveries that may eventually lead to financial gains. For example, a simple analysis may reveal calling towers that are used very frequently, yet are associated with dropped calls and degraded service. Such towers should be given a higher priority for maintenance, before customers within its range decide to change providers as a result of poor service. As another example, users who receive an extraordinarily large number of calls may be targeted for a deeper investigation, to see if the number is being used as a calling center or for some other inappropriate purpose.

Beyond looking for outliers or correlations in a dataset, advanced data analytics are also utilized to find more sophisticated patterns to answer more challenging questions. In this section, we present novel propagation models used in such advanced analytics that predicts the likelihood that a customer will soon *churn*, or move to a different service provider and identifies fraudulent activity in a calling network. Churning is a significant problem for service providers because, in today's society where nearly everyone has a mobile phone, it has become very expensive to attract customers who do not yet have a phone to join their service. Furthermore, today's users are more informed about the kind of devices, the quality of the service, and the perks offered by the providers. Such providers must thus devote a significant amount of effort towards customer *retention*, rather than *acquisition*. Fraudulent activity in a calling network relates to voice-related security threats where users may reveal sensitive or private information through social engineering techniques and by calling international phone numbers. These calls carry a financial cost to both the subscribers and the service provider.

11.5.1 Churn Prediction: Sender-Centric Energy Propagation

The decision to drop a service provider is based not only on a user's own satisfaction with the service, but may also be the result of social pressures from friends, family, and other close contacts who have already decided to churn. Researchers have thus turned to *energy propagation* models across the calling network of a mobile phone provider, where energy refers to information that may persuade another user to churn. In this model, users marked to have churned during a month is seeded with an amount of

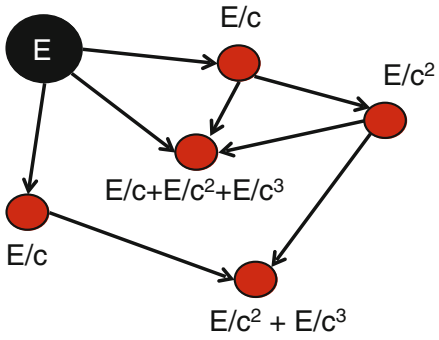

Fig. 11.6 Generic illustration of sender-centric energy propagation

energy E. These churners divide this energy into smaller portions and disseminates it across all of their connections. Users who receives portions of this energy then replicates it, divides it into even smaller portions, and spreads the energy across its contacts. This process of accumulating, dividing, and spreading energy repeats until the fraction of energy received at any user drops below some threshold t. Figure 11.6 illustrates this spreading process. The churner (black node) distributes E/c energy to its three contacts, where c is some positive constant. These three contacts store this energy and then replicate a fraction E/c^2 of it to be sent to each of its own contacts. The total energy accumulated by a user may thus represent the likelihood that she will soon churn from the service provider.

Rather than having every user propagate a constant fraction $1/c$ of its energy to others, we define a *transfer function* $F(c)$ that returns what proportion of stored energy is transferred to each of a user's contacts. This transfer function is defined by the *sender* of the influence, putting them in control of how much energy each recipient will be exposed to. Because the receivers have no choice but to accept the energy it receives and pass it along, we refer to this energy propagation model as being **sender-centric**. Sender-centric propagation models may differ in the way senders choose what contacts to receive, and by how $F(c)$ is defined.

Dasgupta et al. proposed the following sender-centric energy propagation model for churn prediction [11]. Consider a diffusion process where at each time step t there is a set of active users X whose members $x \in X$ have energy $E(x, t)$. At time step $t+1$, every active user in X transfers a fraction of its energy to all of their neighbors y. The fraction of energy sent is a function of two parameters: the spreading factor d and transfer function F. d is a constant that lets the modeler decide by how far the energy propagation should spread. Low values d keep the process very local, while high values of d lets energy spread far away from the churner. F should be designed in a way that reflects the relative 'strength' a connection to one contact is over another, so that more energy is transferred over stronger connections. For example, information shared by a good friend who one has strong connections to will be given higher consideration. If W_{xy} is the strength of a connection from x to y, F may be defined as:

$$F = \frac{W(x, y)}{\sum_{\{(x,s)|s \in N(x)\}} W(x, s)} \tag{11.44}$$

The set of active users at time $t + 1$ is then given by the set of nodes who received energy. The energy propagation process terminates at time t^* if no new nodes are exposed at time t^* or if the amount of energy any node is exposed to falls below a threshold value E_T.

11.5.1.1 Model Application

Dasgupta et al. use the above sender-centric model to predict churners in a mobile call graph [11]. They define connection strength as $W_{xy} = 2/(1 + e^{-c_{xy}}) - 1$ where c_{xy} is the total number of calls placed from user x to y. They then select a threshold energy value T_c, where any user on the network that collects more than T_c energy is predict to become a churner. They investigate the fraction of all churners correctly caught as T_c decreases to include a larger fraction of users on the network. They find that the set of users having the 10 % largest amounts of energy contain approximately 45 % of all churners in a given month. From the perspective of a mobile phone service provider this is a strong result. For example, the provider can invest in a marketing campaign that targets just 10 % of its subscribers with discounts, in an attempt to prevent almost half of all potential churners from switching service providers. By comparison, the 10 % most probable churners labeled by a decision-tree classifier that uses features about the frequency a user utilizes her mobile phone service and her connectivity contains only approximately 40 % of all churners.

11.5.2 Churn Prediction: Receiver-Centric Energy Propagation

In a sender-centric energy propagation model, the transfer function $F(c)$ is defined as a function of some features about the sender of information. However, one may hold the philosophic belief that it is the receiver of information, rather than the sender, who ultimately decides the degree to which she becomes influenced. This idea gives rise to an alternative class of energy propagation models that are *receiver-centric*. The rules that govern a receiver-centric propagation process may be summarized as follows [40]:

1. A user who receives energy by a neighbor will decide what proportion should be retained. This retention should be proportional to the strength of the relationship between the receiver and the sender.
2. A user only retains energy originating from a churner once.
3. However, users retain energy many times if the energies originate from different sources.

4. When a user receives some amount of influence by a neighbor, she chooses the proportion to be retained. She subsequently replicates and transmits this proportion to every one of her neighbors.

The first rule ensures that the total influence retained by a receiver will be grounded in the relationship held between the receiver and sender. The second rule captures the idea that, if a user is exposed to energy from the same source but at different iterations of the propagation, she will only retain energy from the first exposure. Intuitively, multiple exposures of energy originating from the same source would contain same information, which the receiver already considered during her first exposure. The information or influence contained in energy sent from distinct sources, however, is unique. Hence, in the third rule, a receiver is allowed to retain energy multiple times if the source of the energy is distinct. Finally, the receiver will transmit a copy of all energy she retains to all of her contacts. Her contacts will then independently decide how much energy they should retain.

Phadke et al. introduce a receiver-centric model for predicting churners in a mobile phone network [40]. They define a strength for the relationship between users X and Y using a vector of calling attributes $(x_1, \ldots, x_n)$. Each attribute x_i is normalized by dividing it by $|x_i|$, where $|x_i| = \sqrt{\sum_{k=1}^{d} x_{ik}^2}$ so that they are of unit length. For a relationship k, let $k = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n$ be the weighted sum of its normalized attributes. The strength of the relationship between X and Y can be defined by any monotonically increasing function of k; the authors use $W_{XY}(k) = 1 - e^{-k/\varepsilon^2}$. This exponential function is based on the idea that when a strong relationship is established between two users, there is a higher likelihood that the information or influence within the energy passed along that connection will be retained by the receiver. ε is a tunable parameter that controls the degree to which the strength of a relationship is affected by the magnitude of it's attributes.

The model computes the total amount of energy received by a user in an iterative process. It begins with the passing of E energy from every node that churned in the previous month to all of its neighbors. Let N_i be the set of neighbors of node i. A neighbor j of a churner i will choose to retain

$$E_j = \frac{W_{ij}}{\mathbf{W_j}} E_i \tag{11.45}$$

where $\mathbf{W_j}$ is the sum of the strength of all relationships j is a part of and E_i is the energy contained by churner i. These neighbors will then pass E_j units to its neighbors, and so forth, until the number of iterations exceeds a threshold value (in their study, they terminate the process after three iterations). After the process terminates, each receiver adds together all of the energy it received.

11.5.2.1 Model Application

Phadke et al. apply the receiver-centric model to a dataset of calls placed between over half a million users and churners during a two month period. For a single month of data, the authors compute the strength of each connection using the call time, number of calls made, and neighborhood overlap as relationship features. They tuned the weights α_i empirically in order to maximize the predictive accuracy of the propagation model. They then consider a boosted decision tree ensemble classifier that uses the amount of energy retrained along with features such as whether a contract has ended, the number of days a user is connected, the number of calls made to churners, and the charged rate for making phone calls to assign each users a probability that they will churn in the subsequent month. They find that without the energy feature, the classifier finds 35 % of all future churners among the top 10 % most likely users predicted to churn. By adding the energy accumulated, this percentage rises to approximately 40 %. In summary, they find the receiver-centric energy propagation model to be a viable alternative to a sender-centric model.

11.5.3 Isolating Fraudulent Activity: Markov Clustering Algorithm

Jiang et al. [19] present a method for identifying fraudulent activity performed over voice calls in a cellular network by analyzing the structure of a calling network. Their method is rooted in the following features about fraudulent activity on mobile phone networks: (i) callers on a phone network seeking to commit fraud tend to contact a large number of people and will attract more victims to call fraudulent numbers compared to a typical user of the phone network; and (ii) fraudsters may utilize many international phone numbers at once to distribute their scheme, which lets them increase the number of victims that can be reached. This activity may be represented by observing the same set of domestic users (victims) who all call the same set of foreign (fraudulent) numbers.

These two features suggest that fraudulent activity may be characterized by finding *community structures* containing large numbers of international calls to the same collection of phone numbers. To find these communities, the method uses the Markov Clustering Algorithm (MCL). This algorithm finds communities by iterating over two steps: network *expansion* and *inflation*. At iteration i, the expansion step takes the square of the adjacency matrix of the network to simulate the probability of random walks of length $i + 1$ that start and terminate at every user in the call graph. In the inflation step, the elements of the squared adjacency matrix are raised to a power β, and then the matrix is scaled diagonally so that the resulting adjacency matrix is Markovian. In essence, the inflation step modifies the probabilities associated with random walks in a way that favors more probable walks. As the process repeats, matrix entries corresponding to links in low probability walks will converge to zero,

so the converged adjacency matrix will only contain connections in high probability walks. The connected components of this converged matrix correspond to community structures.

To find communities that contain fraudsters, the method looks for 2-by-2 bi-partite cliques from domestic to international numbers. These 2-by-2 bi-partite cliques are the smallest structural unit that corresponds to fraudulent activity, where a set of victims who do not know each other both call the same two fraudulent numbers. The method filters out all communities that do not exhibit at least α bi-partite cliques of any size that have at least γ victims.

11.5.3.1 Model Application

Jiang et al. use the MCL-based method to analyze a dataset of all international voice calls made within the voice network of a major service provider [19]. They take two sources of user reports to build a ground truth list of fraudulent calls, referred to as an international revenue share fraud (IRSF) list: (i) numbers reported by customers to the provider's customer care center; and (ii) a list of phone numbers tied to customer complaints that were posted online in blogs, social media, and forums [1, 50]. They run the MCL detection algorithm on different months of data (Jan–May 2011) to study the expected lag that will occur between when fraudulent activity occurs and when it will be reported in the IRSF or online list of fraud numbers. They choose $\alpha = 5$ and $\gamma = 10$ after observing that these settings filter out over 98 % of the subgraphs while capturing over 90 % of all communities that exhibit fraud. They compare the numbers in these fraud communities against a list of over 24,000 numbers fraudulent numbers covered in the IRSF lists. They find that the extracted communities only contain 11 % of the numbers in the list. However, these 11 % of numbers attract phone calls from 85 % of all victims, and are the root cause of 78 % of all fraudulent calls in the network. Furthermore, when the authors exclude dormant numbers in the IRSF list (numbers not yet utilized or advertised by fraudsters), the detection rate increases from 11 to over 50 %.

The authors also evaluate whether the MCL algorithm can be used to identify fraudsters early, before they are reported or recorded on an IRSF list. For all fraud numbers contained in the communities extracted, the gap between the month it was extracted from in the data and the month it was added to the IRSF list is compared. For more than 80 % of the fraud numbers, the detection method precedes the user reports and in more than 60 % of these cases, the fraud numbers are discovered at least one month sooner than when a report is shared by a user.

11.5.3.2 Summary of Findings

The models presented in this chapter found a number of important characteristics and new findings about mobile phone communication networks. We summarize these findings next.

- **Diffusion processes are governed by heavy-tailed distributions**. The distributions of how long information propagates between two users, and the frequency of these propagations, are characterized by mixtures of Lognormal distributions.
- **Physical co-location is strongly correlated with the formation of future connections** Users that propagate information between each other are likely to be co-located for brief periods of time. Whether or not two users exist in the same location strongly predicts whether they will form new connections in the future.
- **Short-lived information over calling networks does not diffuse widely**. The total number of others that receive short-lived information is strongly correlated with the in- and out-degree distribution of the users participating in the diffusion process. Propagations of short-lived information are generally limited to a very local level and do not spread far and wide across a calling network.
- **Epidemiological models are a flexible tool to understand local-level interactions and the spreading of malware**. Epidemiological models have been used to successfully model the dynamics of malware that spreads at local levels. Different kinds of models can incorporate specific properties of mobile devices, including the range of their transmissions and energy constraints. SIP-based models become less accurate if transmissions can only be performed devices are within very close proximity. SIS-based models may be used in scenarios where devices can never become immunized. SIDR-based models work under scenarios where devices can be killed or disabled by malware. To maximize damage, malware should wait for infections to spread before killing devices. As the recovery rates of devices increase, the total damage of a malware outbreak drops exponentially.
- **Energy propagation models can help identify future churners**. Irrespective of whether a modeler uses a sender-centric or receiver-centric propagation model, we can identify a large proportion of future churners by the total energy or influence they accumulate from past churners. Both sender- and receiver- centric propagation models offer promising results.
- **Finding user communities with bi-partite cliques can identify fraudulent activity**. Bi-partite cliques may correspond to users who send calls to the same subset of fraudulent phone numbers on the network. 80 % of the communities found through a Markov clustering algorithm containing such bi-partite cliques include fraudulent numbers not yet been reported by users.

11.6 Future Research Directions

The state-of-the-art propagation models presented in this chapter represent significant advances in mobile phone data analytics. However, many opportunities remain where researchers may build off of, extend, and use the discoveries made by these methods to propose new kinds of models.We next present a small sampling of these research opportunities.

1. **Marry structure and decisions in the diffusion of information**. The propagation models reviewed in this chapter concentrate on either the *structure* of a

diffusion process or on how individuals *decide* what information should be saved. For example, causality tree models only reason about the probability that certain subsets of a user's connections will be transmitted information within a given time period. Epidemiological models also rely on the structure of the network as users' devices form connections by their spatiotemporal dynamics within a local area. Sender- and receiver-centric energy propagation models, however, simply assume that information spreads widely across all connections. They then concentrate on modeling the process of deciding to retain information, including who makes the decision (sender or receiver) and how that decision is made.

More faithful models of information diffusion should simultaneously consider both structure and decision-making. For example, one should not assume that churners will decide to submit all of their contacts to peer influence. Furthermore, a receiver of short-term information spreading through a causality tree may decide to not propagate the news further if she is disinterested in the information, if her social relationship with the sender is weak, or if she does not believe that her set of contacts would be interested in the information.

2. **Explore the tradeoffs between sender- and receiver-centric propagation**. For the churn prediction problem, both sender- and receiver-centric models have been demonstrated to be similarly successful. Yet these two model types are underpinned by two very different philosophies: one asserts that the person who sends information controls how much the receiver absorbs, while the other believes that the receiver of information individually decides how much they will accept. One kind of model may be more applicable than the other depending on the setting. For example, marketing studies have demonstrated the persuasive effect that a strong advertisement [46] or speaker [44] can have on the amount of information retrained by others. On the other hand, peoples' experiences and knowledge also modulate the amount of information they choose to retain [16]. The settings under which either a sender- or receiver-centric propagation model is more appropriate remains an open question. Hybrid models that integrate both sender and receiver effects may be an effective development.
3. **Build new epidemiological models that operate on other network levels**. Epidemiological models have mostly been applied to local level networks. Although the analogy between the exchange of information among devices that are physically close and the exchange of diseases between people makes applications at the local level intuitive, the spread of information and data need not be restricted by the proximity of devices. For example, there now exist compromised applications that may submit spam messages and fraudulent links to other contacts in a person's address book [2]. Epidemiological models that operate at the contact level may suitably represent the spread of such SMS spam. Furthermore, the spread of rumors and lies across a calling network may be thought of as a systemic spread of mis-information that convinces or (infects) gullible (susceptible) individuals on the network. Thus, an epidemiological model operating at the calling level may characterize the spread of mis-information by accounting for a user's propensity for believing and spreading false information.

4. **Recognize the differences between devices**. Mobile phone devices are built with hardware that supports a variety of technological features. For example, as of 2014, only Android handsets with NFC chips built in are capable of spreading malware to other devices over this medium. Furthermore, devices that either have SMS messaging disabled or cannot support receiving them will not be able to receive information that spreads across this medium. It is thus necessary to consider the heterogeneous mix of devices with varying capabilities within propagation models over mobile phones. Furthermore, differences between devices are not only associated with hardware configurations, but also by their brand. For example, recent intriguing results have found Apple iPhone users to have more connections to others on average, and are more likely to be connected with an iPhone than an Android user [6]. Thus, at the contact level, there may be a higher propensity for information to propagate from one device to another.
5. **Integrate social features**. Ultimately, contact and calling level networks formed out of mobile phone data are *social networks* where the ties users have with many others correspond to offline relationships. Numerous methods in the literature exist to extract the social qualities of such relationships. For example, analysis of ego-network structures can identify users exhibiting egocentric or selfish tendencies [12] as well as those who sport different kinds of social roles [17]. Depending on these roles and tendencies, a user may exhibit different behaviors in a propagation model. For example, egocentric individuals who will speak with everyone simply to be noticed may send new information to all of their contacts, irrespective of whether that information is fact or fiction. Or perhaps users that lie on the periphery of two communities may decide to not let information move from one to another, out of consideration that the other community may be disinterested. We should also consider social features as we assign weights corresponding to the strength, and hence amount of information that propagates, across connections. For example, we know that exceptionally strong and weak social connections prevent a network of mobile phone calls from fragmenting into a large number of disconnected components [13], and are thus critical avenues for information to diffuse widely across the network.

11.7 Concluding Remarks

This chapter presented a collection of recently developed propagation models used for mobile phone data analytics. This collection of models revealed important statistical qualities of information propagation processes over mobile phone networks, were used to model unique propagation phenomena, and utilized in a number of novel applications. Based on the qualities of the models, it identified a number of open opportunities for researchers to develop ever more sophisticated and realistic models of propagation phenomenon within mobile phone networks.

References

1. 800notes: Directory of unknown callers. http://www.whocallsme.com
2. Almeida, T.A., Hidalgo, J.M.G., Yamakami, A.: Contributions to the study of sms spam filtering: new collection and results. In: Proceedings of 11th ACM Symposium on Document Engineering, pp. 259–262. ACM (2011)
3. Anderson, R.M., May, R.M., Anderson, B.: Infectious diseases of humans: dynamics and control, vol. 28. Wiley Online Library (1992)
4. Barron, A., Rissanen, J., Yu, B.: The minimum description length principle in coding and modeling. IEEE Trans. Inf. Theory **44**(6), 2743–2760 (1998)
5. Berlingerio, M., Calabrese, F., Di Lorenzo, G., Nair, R., Pinelli, F., Sbodio, M.L.: Allaboard: a system for exploring urban mobility and optimizing public transport using cellphone data. In: Machine Learning and Knowledge Discovery in Databases, pp. 663–666. Springer (2013)
6. Bjelland, J., Canright, G., Engo-Monsen, K., Sundsoy, P.R., Ling, R.S.: A social network study of the apple vs. android smartphone battle. In: Proceedings of International Conference on Advances in Social Networks Analysis and Mining, pp. 983–987. IEEE Computer Society (2012)
7. Candia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G., Barabási, A.L.: Uncovering individual and collective human dynamics from mobile phone records. J. Phys. A: Math. Theor. **41**, 11 pp (2008)
8. Chen, W., Wang, Y., Yang, S.: Efficient influence maximization in social networks. In: Proceedings of 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 199–208. ACM (2009)
9. Chien, E.: Security response: Symbos. mabir. Techical report. Symantec Corporation (2005)
10. Corporation, I.: Global business security index report. Technical report. IBM (2004)
11. Dasgupta, K., Singh, R., Viswanathan, B., Chakraborty, D., Mukherjea, S., Nanavati, A.: Social ties and their relevance to churn in mobile telecom networks. In: Proceedings of 11th ACM International Conference on Extending Database Technology (2008)
12. Doran, D., Alhazmi, H., Gokhale, S.: Triads, transitivity, and social effects in user interactions on facebook. In: Proceedings of IEEE International Conference on Computational Aspects of Social Networks, pp. 68–73 (2013)
13. Doran, D., Mendiratta, V., Phadke, C., Uzunalioglu, H.: The importance of outlier relationships in mobile call graphs. In: Proceedings of International Conference on Machine Learning and Applications, pp. 24–29 (2012)
14. Eagle, N., Pentland, A.S., Lazer, D.: Inferring friendship network structure by using mobile phone data. Proc. Natl. Acad. Sci. **106**(36), 15274–15278 (2009)
15. Ferrie, P., Szor, P., Stanev, R., Mouritzen, R.: Security response: SymbOS. Symantic Corporation, Cabir. Technical repot (2004)
16. Fessenden-Raden, J., Fitchen, J.M., Heath, J.S.: Providing risk information in communities: Factors influencing what is heard and accepted. Sci. Technol. Hum. Values **12**, 94–101 (1987)
17. Gleave, E., Welser, H.T., Lento, T.M., Smith, M.A.: A conceptual and operational definition of 'social role' in online community. In: 42nd Hawaii International Conference on System Sciences, pp. 1–11 (2009)
18. Groenevelt, R., Nain, P., Koole, G.: The message delay in mobile ad hoc networks. Perform. Eval. **62**(1), 210–228 (2005)
19. Jiang, N., Jin, Y., Skudlark, A., Hsu, W.L., Jacobson, G., Prakasam, S., Zhang, Z.L.: Isolating and analyzing fraud activities in a large cellular network via voice call graph analysis. In: Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, pp. 253–266. ACM (2012)
20. Kaspersky: Kaspersky security bulletin malware evolution. Kaspersky Security Bulletin Malware Evolution 2011
21. Kempe, D., Kleinberg, J., Tardos, E.: Maximizing the spread of influence through a social network. In: Proceedings of 9th ACM International Conference on Knowledge Discovery and Data Mining, pp. 137–146 (2003)

22. Kephart, J.O., White, S.R.: Directed-graph epidemiological models of computer viruses. In: Proceedings of IEEE Computer Society Symposium on Research in Security and Privacy, pp. 343–359. IEEE (1991)
23. Kermack, W.O., McKendrick, A.G.: Contributions to the mathematical theory of epidemics. part i. Proc. Roy. Soc. Lond. Ser. A **115**(5), 700–721 (1927)
24. Kermack, W.O., McKendrick, A.G.: Contributions to the mathematical theory of epidemics. ii. the problem of endemicity. Proc. Roy. Soc. Lond. Ser. A **138**(834), 55–83 (1932)
25. Khouzani, M., Sarkar, S., Altman, E.: Maximum damage malware attack in mobile wireless networks. IEEE/ACM Trans. Netw. **20**(5), 1347–1360 (2012)
26. Kim, H., Zang, H., Ma, X.: Analyzing and modeling temporal patterns of human contacts in cellular networks. In: Proceedings of 22nd IEEE International Conference on Computer Communications and Networks, pp. 1–7 (2013)
27. Kitagawa, G., Gersch, W.: Smoothness Priors Analysis of Time Series, vol. 116. Springer (1996)
28. Krings, G., Calabrese, F., Ratti, C., Blondel, V.D.: Urban gravity: a model for inter-city telecommunication flows. J. Stat. Mech.: Theory Exp. L07003 (2009)
29. Król, D.: Propagation phenomenon in complex networks: theory and practice. New Gener. Comput. **32**(3–4), 187–192 (2014)
30. Kurtz, T.G.: Solutions of ordinary differential equations as limits of pure jump markov processes. J. Appl. Probab. **7**(1), 49–58 (1970)
31. Lambiotte, R., Blondel, V.D., de Kerchove, C., Huens, E., Prieur, C., Smoreda, Z., Van Dooren, P.: Geographical dispersal of mobile communication networks. Phys. A Stat. Mech. Appl. **387**(21), 5317–5325 (2008)
32. Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., Glance, N.: Cost-effective outbreak detection in networks. In: Proceedings of 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 420–429. ACM (2007)
33. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., Tomkins, A.: Geographic routing in social networks. Proc. Natl. Acad. Sci. USA **102**(33), 11623–11628 (2005)
34. Mao, H., Shuai, X., Ahn, Y.Y., Bollen, J.: Mobile communications reveal the regional economy in cote d'ivoire. In: Proceedings of International Conference on Analysis of Mobile Phone Datasets and Networks D4D Book, pp. 1–18 (2013)
35. Mickens, J.W., Noble, B.D.: Modeling epidemic spreading in mobile environments. In: Proceedings of the 4th ACM Workshop on Wireless Security, pp. 77–86. ACM (2005)
36. Montoliu, R., Gatica-Perez, D.: Discovering human places of interest from multimodal mobile phone data. In: Proceedings of 9th International Conference on Mobile and Ubiquitous Multimedia, pp. 12:1–12:10. ACM (2010)
37. Pan, W., Aharony, N., Pentland, A.: Composite social network for predicting mobile apps installation. In: Proceedings of AAAI Conference on Artificial Intelligence, pp. 821–827 (2011)
38. Peng, S., Yu, S., Yang, A.: Smartphone malware and its propagation modeling: a survey. IEEE Commun. Surv. Tutor. **16**(2), 925–941 (2014)
39. Peruani, F., Tabourier, L.: Directedness of information flow in mobile phone communication networks. PloS One **6**(12), e28,860 (2011)
40. Phadke, C., Mendiratta, V., Uzunalioglu, H., Doran, D.: Prediction of subscriber churn using social network analysis. Bell Labs Tech. J. **17**(4), 63–75 (2013)
41. Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., Ratti, C.: Activity-aware map: identifying human daily activity pattern using mobile phone data. In: Human Behavior Understanding, pp. 14–25. Springer, Berlin (2010)
42. Rhodes, C.J., Nekovee, M.: The opportunistic transmission of wireless worms between mobile devices. Phys. A Stat. Mech. Appl. **387**(27), 6837–6844 (2008)
43. Rivera, M.T., Soderstrom, S.B., Uzzi, B.: Dynamics of dyads in social networks: assortative, relational, and proximity mechanisms. Annu. Rev. Sociol. **36**, 91–115 (2010)
44. Sellnow, T.L., Ziegelmueller, G.: The persuasive speaking contest: an analysis of twenty years of change. Natl. Forensic J. **6**(2), 75–87 (1988)
45. Systems, JJuniper: Mobile Threats Report. In: Technical report (2011)

46. Taillard, M.O.: Persuasive communication: the case of marketing. Working Papers in Linguistics, vol. 12, pp. 145–174 (2000)
47. Trivedi, K.S.: Probability and Statistics with Reliability, Queueing, and Computer Science Applications, 2nd edn. Wiley, New York (2002)
48. Wang, D., Pedreschi, D., Song, C., Giannotti, F., Barabasi, A.L.: Human mobility, social ties, and link prediction. In: Proceedings of 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1100–1108. ACM (2011)
49. Wang, Y., Cong, G., Song, G., Xie, K.: Community-based greedy algorithm for mining top-k influential nodes in mobile social networks. In: Proceedings of 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1039–1048. ACM (2010)
50. WhoCallsMe: Reverse phone number lookup. http://www.whocallsme.com
51. Zhang, W., Li, Z., Hu, Y., Xia, W.: Cluster features of bluetooth mobile phone virus and research on strategies of control and prevention. In: Proceedings of International Conference on Computational Intelligence and Security, pp. 474–477. IEEE (2010)