

Identifying Significant Facilitators of Dark Network Evolution

Daning Hu

*Department of Management Information Systems, University of Arizona, Tucson, AZ 85721.
E-mail: hud@email.arizona.edu*

Siddharth Kaza

*Department of Computer and Information Sciences, Towson University, Towson, MD 21252.
E-mail: skaza@towson.edu*

Hsinchun Chen

*Department of Management Information Systems, University of Arizona, Tucson, AZ 85721.
E-mail: hchen@eller.arizona.edu*

Social networks evolve over time with the addition and removal of nodes and links to survive and thrive in their environments. Previous studies have shown that the link-formation process in such networks is influenced by a set of facilitators. However, there have been few empirical evaluations to determine the important facilitators. In a research partnership with law enforcement agencies, we used dynamic social-network analysis methods to examine several plausible facilitators of co-offending relationships in a large-scale narcotics network consisting of individuals and vehicles. Multivariate Cox regression and a two-proportion z-test on cyclic and focal closures of the network showed that mutual acquaintance and vehicle affiliations were significant facilitators for the network under study. We also found that homophily with respect to age, race, and gender were not good predictors of future link formation in these networks. Moreover, we examined the social causes and policy implications for the significance and insignificance of various facilitators including common jails on future co-offending. These findings provide important insights into the link-formation processes and the resilience of social networks. In addition, they can be used to aid in the prediction of future links. The methods described can also help in understanding the driving forces behind the formation and evolution of social networks facilitated by mobile and Web technologies.

Introduction

The notion of social networks and the methods of social-network analysis (SNA) have received great interest from the

information-science community in recent years. The use of computer systems in the form of social networking and bookmarking Web sites (Thelwall, 2008), cell phones and other mobile smart devices, and intraorganization virtual collaboration methods (Haythornthwaite, 2006) have provided rich data sources for studying various large-scale social networks. There is a great need to understand the impact of these social networks and the extent to which they are facilitated by various social and technological factors. In this study, we focus on the facilitators that lead to link formation in social networks. These facilitators may be attributes of the individuals involved (e.g., age, gender), their network contexts (e.g., mutual acquaintances), and other kinds of nodes (e.g., use of vehicles, use of communication technology). We use an example of a dark network of criminals in this study and utilize novel methods to study its facilitators.

Illegal activities such as drug trafficking, money laundering, and terrorist attacks are usually conducted by individuals operating in networks (Chen, 2006). These illegal and covert social networks are referred to as “dark networks” (Raab & Milward, 2003). Like “bright” organizational networks, dark networks also serve as communication, collaboration, and coordination mediums to better achieve goals. The networks consist of a set of nodes/actors (e.g., criminals, terrorists) and links/relationships among them (e.g., crimes, terrorist operations, kinship). For instance, a network may consist of criminals who committed a bank robbery as nodes and their co-offending relationships in that robbery as links.

Networks evolve over time with the addition and removal of nodes and links. Often, the evolution is influenced by external factors. For instance, a terrorist network may grow in a conducive environment like a failed nation-state with

Received July 3, 2008; revised November 3, 2008; accepted November 3, 2008

© 2009 ASIS&T • Published online 8 January 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.21008

ample resources. A social-networking-site friend network might grow with the addition of an instant messaging application. Previous studies in sociology (Kossinets & Watts, 2006; McPherson, Smith-Lovin, & Cook, 2001) have shown that the link-formation process in such social networks is influenced by a set of facilitators. These facilitators may be individual attributes like age, race, and gender (Feld, 1982; McPherson et al., 2001; Thelwall, 2008; Reiss, 1986) of the nodes/actors in the network, or shared affiliations between actors, like kinship and mutual acquaintances (Backstrom, Huttenlocher, Kleinberg, & Lan, 2006; Kossinets & Watts, 2006).

Previous studies on the facilitators of dark-network evolution have primarily viewed them from a qualitative standpoint with few empirical evaluations (Coles, 2001; McPherson et al., 2001; Milward & Raab, 2006; Sarnecki, 2001). In this study, dynamic social-network analysis (SNA) methods are used to examine several plausible link-formation facilitators in a large real-world narcotics network. We use multivariate survival analysis using Cox regression and a two-proportion z-test on the network cyclic and focal closure to identify significant facilitators. This study aims to answer the following questions:

- What are the facilitators of the link-formation process in an evolving social network?
- How can we quantitatively identify the significant facilitators that influence the link-formation process in a social network?

The knowledge of significant facilitators can be used to intervene in the formation of networks and focus efforts to encourage or discourage the formation of future links. The methodology used can also be generalized to study the effects of various facilitators on networks of individuals and machines. The remainder of the paper is organized as follows: The next section reviews previous literature on the study of networks and dynamic SNA methods. The subsequent section introduces the testbed for this study. The research design is described, then experimental results are presented and discussed. Finally, we conclude and suggest directions for future work.

Research Background

SNA has been used to study various real-world networks including dark networks (Krebs, 2001; Raab & Milward, 2003; Sarnecki, 2001; Yang & Li, 2007). There are two main kinds of SNA studies. One focuses on the static topology of social networks where the structural properties of the nodes and links are examined to describe and explain how network topologies affect the functions and behaviors of complex systems (Albert & Barabasi, 2002). Such SNA studies have mainly focused on two types of social networks: (a) communication networks such as Internet (Albert, Jeong, & Barabasi, 1999), e-mail (Newman, Forrest, & Balthrop, 2002) and phone (Abello, Pardalos, & Resende, 1999) networks; and (b) collaboration networks such as co-authorship networks (Barabasi, Jeong, Neda, et al., 2002; Newman, 2001) and movie-actor networks (Watts & Strogatz, 1998).

However, these studies have largely ignored dynamic network processes such as link formation.

Another line of SNA research has thrived by studying various dynamic network processes and the mechanisms and determinants behind those processes. Such dynamic social-network analyses mainly use statistical methods to model different network processes. These models are then tested to account for the structural changes of network topologies. The network processes studied include formations of friendship links (Leenders, 1996; Snijders, Steglich, & Schweinberger, 2007), collaboration links (Lomi & Pattison, 2006; Kaza & Chen, 2009; Nerkar & Paruchuri, 2005; Powell, White, Koput, & Owen-Smith, 2005), and communication links such as e-mails (Kossinets & Watts, 2006) and phone calls (Palla, Barabasi, & Vicsek, 2007).

The main issues with dynamic SNA include network-recovery techniques, appropriate network measures, and statistical analysis methods for analyzing evolving networks. These issues are discussed in the following subsections.

Dynamic Social-Network Analysis Methods

This section reviews the three major issues of dynamic SNA (Kossinets & Watts, 2006): network recovery, network measurement, and statistical analysis. There are various techniques to address these issues. The drawbacks and advantages of these techniques are also discussed.

Network recovery. Recovery is the process by which multiple instantaneous network representations are recovered from longitudinal data to model an evolving network. Recovery techniques can be classified into discrete and continuous techniques according to their analytically different perceptions of time (Moody, McFarland, & Bender-deMoll, 2005). Discrete network-recovery techniques take multiple cross-sectional snapshots of the network at fixed points in time. However, these snapshots usually do not capture the processes that lead to network evolution. Most previous dynamic SNA studies (Leenders, 1996; Leskovec, Kleinberg, & Faloutsos, 2005; Xu, Marshall, Kaza, & Chen, 2004) used the discrete technique due to the lack of methods for continuous network recovery and the high computational complexity for the few methods that existed.

Some recent dynamic SNA studies (Kossinets & Watts, 2006; Moody et al., 2005) have used continuous recovery techniques to extract multiple instantaneous networks from longitudinal data. These instantaneous networks account for the processes that lead to changes in link structure. In order to model the evolution in an efficient manner, Moody et al. proposed a method to aggregate sequential events into larger time units. Based on this idea, Kossinets and Watts (2006) proposed a sliding window filter to smooth the link-formation process. The time-span of the sliding window (called the “relevancy horizon”) defines which past events are relevant to the current state of the network. In other words, the sliding window is the amount of time in which the formation of the network stabilizes and thus past events

can be considered as evolutionary events leading to the current state.

In order to determine the time interval between two consecutive instantaneous networks (called the “sampling period”), Kossinets and Watts (2006) suggested applying the Nyquist sampling theorem (Oppenheim & Schaffer, 1989) to the maximum rate of link formation. This theorem states that the continuous link-formation process can be captured by a sampling period $\delta \leq c/(2 \cdot f_{max})$, where c is the number of independent clusters in the network and f_{max} is the maximum frequency of link formation.

The sliding window filter combined with the sampling period provides a novel and efficient method to recover the network. These techniques have a clear advantage over the random sampling periods used by previous studies that do not capture actual link-formation processes.

Network measurement. Most empirical studies on longitudinal data plot descriptive measures over time to describe network changes. There are three main categories of descriptive measures used in dynamic SNA: deterministic measures, probabilistic measures, and temporal measures.

Deterministic measures. Deterministic measures include statistics like network size, measures based on the number of links (like degree, average degree), closeness (Bavelas, 1950), and betweenness (Freeman, 1977). There have been various studies that have used these measures for static network analysis. Albert and Barabasi (2002) provide a comprehensive review. A few dynamic SNA studies (Barabasi, Jeong, Zéda, et al., 2002; Leskovec et al., 2005) have also used these methods to measure network evolution. Barabasi, Jeong, Zéda, et al. (2002) studied the changes in the number of links, nodes, and average degree in a scientific co-authorship network. They found the network was growing since all three measures were constantly increasing during that time period. Leskovec et al. (2005) discovered that several evolving networks became denser over time by identifying their increasing average degrees. In the dark-network problem domain, Xu et al. (2004) studied an evolving criminal network by analyzing and visualizing changes in degree, betweenness, and closeness. However, these studies determined the measures over discrete (and sometimes arbitrary) instances of time. Thus, the studies may not model the actual evolutionary changes in the network.

Probabilistic measures. The two most commonly used probabilistic measures are degree distribution and clustering coefficient. Nodes in a network have different number of links connecting them. Degree distribution is the probability $P(k)$ that a randomly selected node has exactly k links (Albert & Barabasi, 2002). The degree distribution is used to classify networks to different topologies like random (Erdos & Renyi, 1960; Newman, 2001) and scale-free (Barabasi & Albert, 1999). The degree distribution can also be used to explain dynamic processes like growth and preferential attachment in scale-free networks (Barabasi & Albert).

The clustering coefficient is the probability that two nodes with a common neighbor also link to each other (Newman, Barabasi, & Watts, 2006). The measure was introduced to determine the small-world nature of a network. In a dynamic analysis study of scientific collaborations (Barabasi, Jeong, Zéda, et al., 2002), the clustering coefficient was found to decay over time, signifying that the network became less interconnected.

Temporal measures. Temporal measures have been developed to describe continuous network processes by taking the time variable into consideration. The triadic closure (Rapoport, 1953) is defined as the empirical probability that two unconnected nodes (at time t) with a common neighbor will form a new link at time $t + \delta$. Intuitively, the triadic closure is a clustering coefficient within a time period $(t, t + \delta)$. A general form of the triadic closure is called cyclic closure (Kossinets & Watts, 2006), where the two nodes (i and j) may not share a neighbor but can be a certain geodesic distance ($d_{ij} > 2$) apart. Another temporal measure known as focal closure (Kossinets & Watts) defines the probability that two previously unconnected nodes that are some distance apart and share one or more affiliations will form a new link. Thus, the focal closure is essentially similar to the cyclic closure, however, in this case the probability is also influenced by the presence of an affiliation. If the focal closure of a network is consistently larger than the cyclic closure then it can be assumed that the link-formation processes in a network are influenced by selected shared affiliations (Kossinets & Watts).

A drawback of the focal closure is that it cannot be used to study the effect of individual attributes (like gender or race). This is because by definition it measures the temporal changes due to shared affiliations like mutual acquaintances or inmates affiliations (shared affiliations are also known as interaction focuses, hence the name “focal” closure). However, individual attributes like race and gender are static over time, thus their effect cannot be measured by focal closure. Even so, these temporal measures provide distinct advantages over deterministic and probabilistic measures. Even though the degree distribution can be used to model growth (using preferential attachment), it cannot be used to quantify the influences of network facilitators. The cyclic/triadic closure subsumes the clustering coefficient and enhances it to add temporal information. In this study, we use focal and cyclic closure for measurement. To the best of our knowledge, these measures have never been used in the dark-networks problem domain.

Statistical analysis. In network analysis, statistical methods are usually used to explain the emergence of network topologies (like random, scale-free, and small-world networks; Albert & Barabasi, 2002); However, in this study we focus on statistical methods that have been used to identify significant link-formation facilitators. A study by Leenders (1996) used a continuous-time Markov model on longitudinal network data where nodes were individuals and links were

friendship between them. The study found that the gender affiliation significantly affected the link formation between children. However, the continuous time model used in this study assumed that only the state of the network at time $t - 1$ affects the current state (at time t). This may not be a valid assumption for most real-world networks.

Snijders (1996, 2001) developed a class of actor-oriented models to explain network evolution. The models are based on the assumption that nodes adjust their positions in the network based on certain parameters. However, this model assumes that the nodes are aware of their positions with respect to the whole network. This assumption may not be true in dark networks since they are covert in nature. In addition, Snijders (2004) also proposed the use of the independent arc model and the reciprocity model to represent different network effects in evolving networks. Together with the actor-oriented model, these three models are all based on the assumption that the observed networks are outcomes of a Markov process evolving in continuous time.

In order to handle networks with missing information, Carley, Dombroski, Tsvetovat, Reminga, and Kamneva (2003) developed the metamatrix approach to combine a set of networks of people, groups, knowledge, resources, events, or tasks to predict behavior. In a recent study, Kossinets and Watts (2006) used Cox regression analysis to identify significant facilitators in a university campus e-mail communication network. They found that the mutual acquaintance and shared class affiliations (among others) had a statistically significant effect on future link formation. A similar survival-analysis approach was also used by Nerkar and Paruchuri (2005) to determine that network centrality of inventors had a statistically significant effect on the intrafirm citation of their patents. The survival-analysis approach lends itself well to the dark-network problem domain since it does not make any assumptions about the underlying network. Kossinets and Watts (2006) also compared the focal and cyclic closure to determine if shared affiliations play a role in link formation. However, they did not use a statistical significance test in the comparison. In this study, we propose to use the two proportion z-test to compare the two measures.

Dark-Network Studies

Two main streams of research can be identified in dark-network literature: the study of link-formation facilitators and the use of statistical methods to measure existing networks.

Link-formation facilitators. Most studies on link-formation facilitators in dark networks have been done in the fields of sociology and criminology. Raab and Milward (2003) studied organizational changes in two real-world dark networks: the Al-Qaeda terrorist network and the Columbian cocaine trafficking network. They found that a set of facilitators motivated individuals to form or deactivate links. A later study by the same group (Milward & Raab, 2006) suggested that prison might be the most effective facilitator for drug trafficking. They contended that individuals who are jailed in close

proximity to each other are likely to form future links in the network (i.e., co-offend in future crimes).

Various criminology studies (Reiss, 1986; Reiss & Farrington, 1991) have suggested that individual attributes like age and race play important roles in co-offending. Such attribute-based homophily has also been suggested by social-network studies in other domains (Louch, 2000; McPherson et al., 2001). A survey on male criminals in London showed that they were more likely to co-offend with individuals of the same age (Reiss & Farrington, 1991). In addition, the same study found that co-offending by criminals of different races was rare. A study on the dynamics of delinquent groups (Warr, 1996) found that similarity in gender had a mixed effect on relationship formation. Male youth offenders generally followed males. Older females were more likely to co-offend with other males, whereas younger females were likely to co-offend with members of the same sex. Proximity of living/working was also considered as a significant factor leading to co-offending (Milward & Raab, 2006; Reiss & Farrington, 1991).

However, most of the above studies used small-scale datasets that usually involved a few hundred criminals over short time periods. Moreover, the findings are based on basic descriptive statistics instead of systematic statistical and SNA methods.

Statistical analysis of dark networks. There have been a few recent studies that have explored the use of SNA methods for dark networks. Most have focused on assigning roles to actors in the network. Sparrow (1991) explored the use of centrality measures to identify key actors in criminal networks. Following this approach, Krebs (2001) used centrality measures to identify the group leader of the September 11th hijackers. Another terrorist-network study calculated the average degree of the Jemaah Islamiyah terrorist network (Koschade, 2006) and uncovered that the 2002 Bali bombing cell had a high density that allowed it to sustain member losses. Xu and Chen (2004) used SNA methods to determine the leader and gatekeeper role for individual nodes and used hierarchical clustering methods to identify subgroups in criminal networks. By primarily concentrating on the node properties, none of these SNA studies have focused on the facilitators of network link formation in dark networks.

Research Testbed

The testbed for this study was consolidated from two related real-world crime-related datasets: police incident reports from Tucson Police Department (TPD) and inmate information from the Arizona Department of Corrections (ADOC).

Tucson Police Department incident reports contain information on 2.03 million individuals and 1.34 million vehicles involved in illegal activity in the Tucson, Arizona metropolitan region (a population of about 1 million) from 1990–2005. This comprehensive dataset is representative of typical crime-incident databases of midsize cities in the United States. The

dataset was used to extract a narcotics network with individuals as nodes and crime incidents as links. Individuals were included as nodes if they were wanted, suspected, or arrested for narcotics or related crimes. We also extracted information on individual attributes like age, race, and gender. Two individuals in the network were connected by a link if they were in the same incident report involving a narcotics crime (such as sale or possession of drugs) or a narcotics-related crime (such as homicide, aggravated assault, or armed robbery). The time of the incident was also extracted, and was used to trace the evolution of the network. Two subsets (2002–2003, 2004–2005) of this narcotics network were used in this study. These subsets provide the most recent and complete snapshots of the narcotics network. They were limited to 2 years since based on domain-expert feedback it was suggested that a criminal cannot be considered as a part of the network if he/she does not commit a crime for more than 2 years. The size of these subsets was comparable or bigger than previous network-analysis studies (Leskovec et al., 2005; McPherson et al., 2001; Newman, 2001; Xu & Chen, 2005; Xu et al., 2004) and also allowed for efficient computational times. Table 1 lists the key statistics for the subset.

The Arizona Department of Corrections dataset contains information (such as names, dates of birth, and inmate housing facility) for 165,540 jailed individuals in Arizona from 1986 to 2006. The 43,220 individuals in this dataset were also found in the Tucson Police Department dataset. This overlap was used to extract the inmate affiliation between individuals in the network subsets. Two individuals were considered to have an inmate affiliation if they were housed in the same facility during the same time period.

TABLE 1. Key statistics of network subsets.

	2002–03	2004–05
Number of individuals	5,076	4,101
Number of links	7,135	5,693

Research Design

Figure 1 shows the research design for identifying significant link-formation facilitators of a dark network. The process consists of four components. The first component, facilitator identification, involves selecting the link-formation facilitators that need to be tested for significance. These facilitators may be selected based on previous studies or theoretical conjectures on dark networks. The second component, network recovery, contains methods to extract instantaneous networks from longitudinal network data. A set of instantaneous networks represents the evolution of the dark network. The third component, network measurement, involves calculating SNA measures like focal and cyclic closure. The last component, statistical analysis, involves identifying the significant facilitators from individual attributes and shared affiliations using multivariate survival analysis and two-proportion z-test. The details of the design are introduced in the following sections.

Facilitator Identification

In this study, the facilitators included three individual attributes and five shared affiliations. These were selected based on previous criminology and sociology studies in this domain. The individual attributes tested were homophily in age (Reiss, 1986; Reiss & Farrington, 1991), race (Reagans, 2005; Reiss & Farrington), and gender (Reiss; Warr, 1996). For the statistical analysis these variables were operationalized as follows:

- Age: The age was set to “1” if the difference in two individuals’ age was less than or equal to 1 year, and “0” otherwise.
- Race: The testbed contained five possible categories of ethnic origin: Caucasian, African American, Hispanic, Asian, and Native Indian. The race was set to “1” if the individuals had the same ethnic origin and “0” otherwise.
- Gender: The gender was set to “1” if the individuals had the same gender and “0” otherwise.

The shared affiliations tested were mutual acquaintance, inmate affiliation, vehicle affiliation, phone affiliation, and

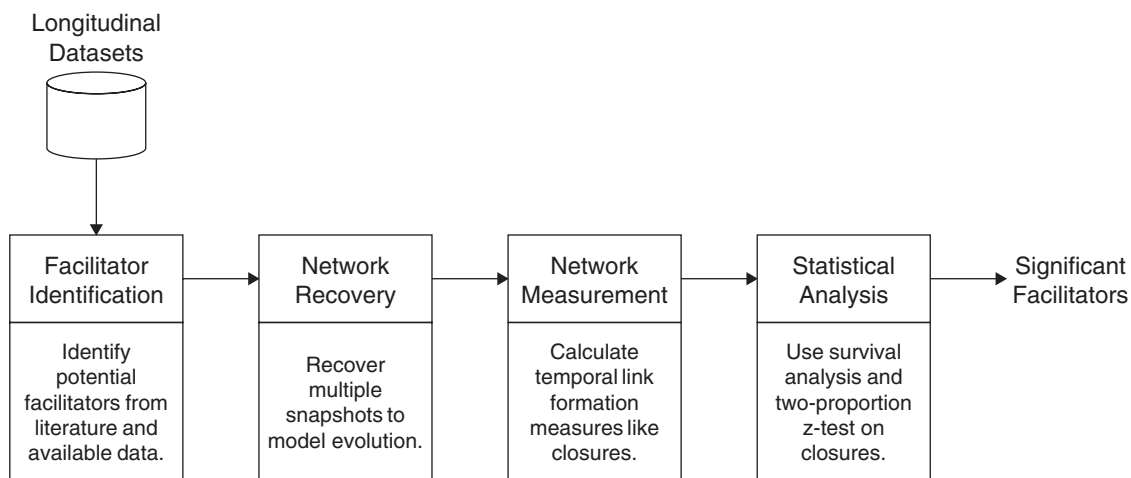


FIG. 1. Research design: identifying significant link-formation facilitators.

residential address affiliation. The first three affiliations were included since they are suggested by previous studies (Coles, 2001; Milward & Raab, 2006). The phone affiliation and residential affiliation were included since they indicated that the individuals worked or lived close to each other. For the statistical analysis these variables were operationalized as follows:

- Mutual acquaintance affiliation: The variable was set to one less than the number of common neighbors two individuals had (Kossinets & Watts, 2006).
- Inmate affiliation was set to “1” if the two individuals were housed in the same prison facility in the same period, and “0” otherwise.
- Vehicle affiliation was set to “1” if the two individuals were found to be associated to the same vehicle in different police reports and “0” otherwise. If two individuals were related to a vehicle in the same police report then this was not considered as a vehicle affiliation since such individuals would be directly linked.
- Phone affiliation was set to “1” if the individuals had the same work or home phone number associated with them and “0” otherwise.
- Residential address affiliation was set to “1” if two individuals live in the same residential grid in Tucson. The Tucson city region is divided into 0.5-square mile administrative grids by the Tucson Police Department. The residential grid can be derived from the address of an individual.

Not all the individuals in the dataset have all the information (e.g., phone or address) associated with them. In case an individual did not have data then the corresponding variable was set to zero. Table 2 shows the number of individuals (in the two network subsets) who had the individual attributes and affiliations listed.

Network Recovery

Recovery is the process by which multiple instantaneous network snapshots are recovered from longitudinal data to model the evolving network. We used the relevancy horizon (τ) and sampling period (δ)—described in the research background section—to recover the network. We used $\tau = 210$ days because the rate of the formation of new links between each pair of nodes stabilizes after approximately 210 days. This was determined using the within-pair response time t_{ij} between two individuals in the network (Kossinets & Watts, 2006). In this dataset, the within-pair response time is defined as the time interval between two subsequent police reports involving the same pair of criminals. Figure 2(a) and 2(b) show the cumulative distribution of within-pair response time for the 2002–2003 and 2004–2005 subset respectively.

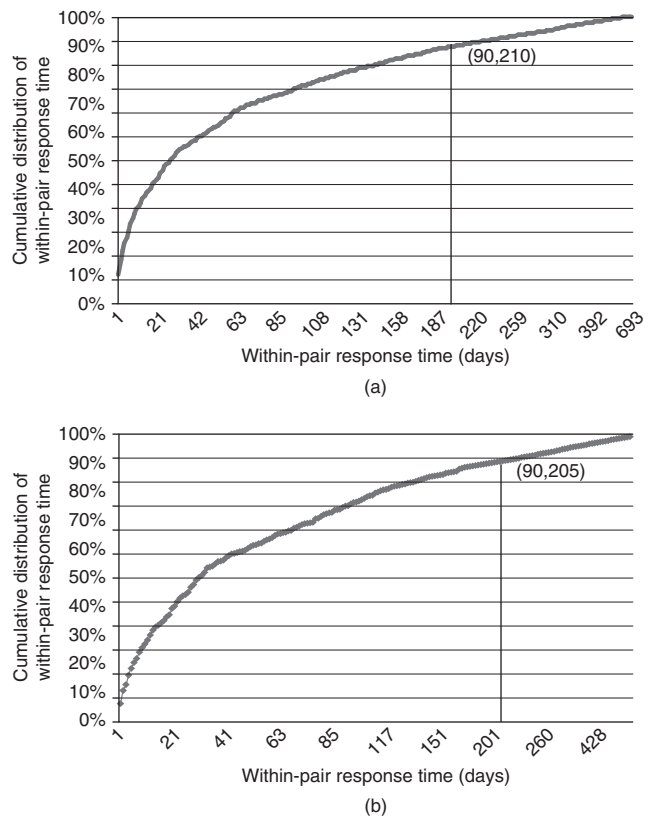


FIG. 2. Cumulative distribution of within-pair response time for the (a) 2002–2003 subset and (b) 2004–2005 subset.

On the x -axis, the within-pair response time (i.e., the number of days between subsequent co-offending) is plotted. The y -axis shows the cumulative distribution of the response time. The graph can be used to determine the time period within which most co-offending takes place. As shown in the figures, 90% of subsequent co-offending happened within $t_{90\%} = 210$ days in the 2002–2003 subset and $t_{90\%} = 205$ days in the 2004–2005 subset.

The sampling period was determined using the Nyquist theorem ($\delta = c/2 \cdot f_{max}$) as described above. For the network under study, c was approximated to be 1,794 and $f_{max} = 19$ links/day, so δ was found to be equal to 47 days. Thus, the evolution of the network was modeled by $(730 - \tau)/\delta = 11$ instantaneous networks recovered from the dataset of 730 days (2 years).

Network Measurement

Once instantaneous networks were recovered from the subset, we calculated the focal and cyclic closures.

TABLE 2. Number of individuals with facilitator information in each subset.

	Age	Race	Gender	Mutual acquaintance	Inmate	Vehicle	Phone	Residential
2002–2003	5,076	5,076	5,076	3,463	3,030	1,829	581	327
2004–2005	4,101	4,101	4,101	2,586	2,346	1,303	434	279

They were calculated using (Kossinets & Watts, 2006):

$$P_{new}(d_{ij}, s_{ij}) = \frac{\sum_{n=1}^{(T-\tau)/\delta} M_{new}(d_{ij}, s_{ij}, n)}{\sum_{n=1}^{(T-\tau)/\delta} M(d_{ij}, s_{ij}, n)}$$

where d_{ij} is the shortest path length between individuals i and j , and s_{ij} is the number of shared affiliations between them. $M(d_{ij}, s_{ij}, n)$ is the number of individual pairs who have s_{ij} shared affiliations and whose shortest path length is d_{ij} for the n th recovered instantaneous network. $M_{new}(d_{ij}, s_{ij}, n)$ is the number of newly formed links that have the same d_{ij} and s_{ij} since the $(n - 1)$ th instantaneous network. $P_{new}(d_{ij}, s_{ij})$ is the probability that a new link will form between any two previously unconnected individuals i and j who are d_{ij} distance apart and have s_{ij} shared affiliations.

The focal closure is $P_{new}(d_{ij}, s_{ij})$ when $s_{ij} > 0$, while the cyclic closure is $P_{new}(d_{ij}, s_{ij})$ when $s_{ij} = 0$. The triadic closure is then represented by $P_{new}(d_{ij} = 2)$. The triadic closure is a special case of the cyclic closure that measures the probability of a new link formation between two previously unconnected individuals whose shortest path length $d_{ij} = 2$.

Statistical Analysis

In this study, two methods were used to identify significant facilitators: (a) a comparative two-proportional z-test on the focal and cyclic closures and (b) a multivariate Cox regression. As mentioned before, a disadvantage of the focal closure is that it can be calculated on affiliations and not individual attributes. Thus, it cannot be used to identify significant attributes. On the other hand, the Cox regression can be used to examine the significance all facilitators (including individual attributes and affiliations).

Two-proportional z-test on focal and cyclic closures. The focal and cyclic closures for each shared affiliation were compared to each other. Kossinets and Watts (2006) suggest that a higher focal closure (as compared to cyclic) between a pair of nodes indicates that an affiliation increases the probability of a link forming between them. We build on their study by incorporating statistical significance to test the difference in the values of focal and cyclic closure. Since the focal and cyclic closures deal with two different populations (one with affiliations and one without), the two-proportion z-test can be used to compare them. The test provides a

greater validity to comparison and helps identify significant affiliations.

Multivariate Cox regression. There are two major types of survival regression models: log-linear and Cox regression models. Using goodness-of-fit tests and checking for proportionality assumptions, we found that the Cox regression model fit our data well. We used Cox regression model of the form (Afifi, Clark, & May, 2003):

$$h(t, x_1, x_2, x_3 \dots) = h_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \dots)$$

where $h(t, x_1, x_2, x_3 \dots)$ is instantaneous hazard—the probability that the event will happen at time t , given that the event has not happened up until time t with the observations of various independent variables ($x_1, x_2, x_3 \dots$). The event in this network is that two previously unconnected nodes with $d_{ij} = 2$ subsequently form a new link. In order to minimize possible correlations among the pairs of nodes with $d_{ij} = 2$, each node was included in only one pair for the analysis. The facilitators described in the previous section were the eight independent variables in the regression. On running the regression, the significance of each of the variables can be ascertained.

Experimental Results and Discussion

Table 3 shows the results of the comparative z-test on the difference of focal closure and cyclic closure for the shared affiliations. We calculated the measures for pairs of nodes with geodesic distance (d_{ij}) of two and three. This was based on domain-expert feedback that indicated that only individuals two or three crime-based links away from a pair are likely to have an effect on their future link formation. In addition, our previous study (Kaza, Xu, Marshall, & Chen, in press) indicated that almost the entire narcotics network in our dataset could be reached within a geodesic distance ranging between four and five. Thus, using $d_{ij} = 4$ would have included almost all the possible node pairs in the network leading to impractically large computation times. The focal closure cannot be calculated for $d_{ij} = 2$, since by definition all pairs of individuals with $d_{ij} = 2$ shared at least one mutual acquaintance. Thus, focal and cyclic closure will be exactly the same for such individuals. In addition, at $d_{ij} = 3$, no pairs of individuals will have mutual acquaintances.

As can be seen from Table 3, the focal closure with the vehicle affiliation was found to be significantly larger than

TABLE 3. Results of two-proportional z-tests on focal and cyclic closures.

Time period	Number of nodes	Geodesic distance	p-value			
			Inmate	Vehicle	Phone	Residential
2002–2003	5,076	2	>0.999	0.012*	>0.999	>0.999
		3	0.641	0.002*	0.058**	>0.999
2004–2005	4,101	2	0.063**	0.004*	0.999	>0.999
		3	0.242	0.022*	>0.999	>0.999

Note. * $p < 0.05$, ** $p < 0.10$.

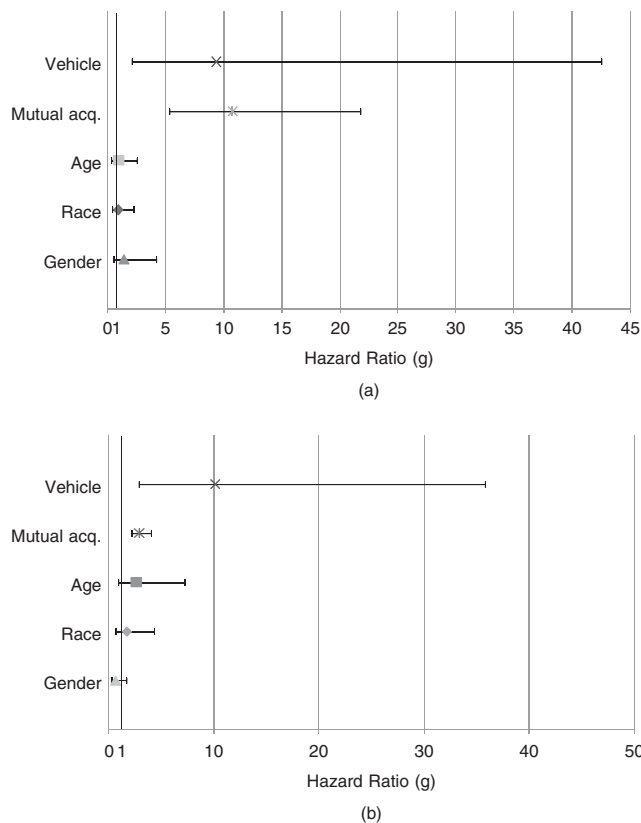


FIG. 3. Results of multivariate survival analysis (Cox regression) of triadic closure for pairs of individuals from (a) 2002–2003 and (b) 2004–2005.

the cyclic closure for $d_{ij} = 2$ and $d_{ij} = 3$ for both subsets. This finding suggests that two previously unconnected criminals who were linked to a vehicle through different crimes are much more likely to co-offend in future crimes than criminals who do not have the vehicle affiliation. The inmate and phone affiliation were found to be mildly significant for one subset, showing some support for the theory suggested by Milward and Raab (2006). We discuss more of the implications of significant and insignificant facilitators below.

Figure 3 shows the hazard ratios and their 95% confidence intervals for the independent variables examined for both network subsets. Each facilitator was represented by an independent variable, and the probability of the triadic closure would increase by a factor of hazard ratio (g) when the corresponding independent variable increases by one unit. The independent variables were considered to be significant

only when the confidence intervals did not include the value 1, since a hazard ratio of 1 indicates that the independent variable has no effect on the dependent variable. Three variables representing inmate affiliation, phone affiliation, and residential affiliation were dropped from the Cox model due to collinearity. The confidence intervals in Figure 3 show that the vehicle and mutual acquaintance affiliation were found to be significant with hazard ratio ranges not including the value 1. The p -values for each of the variables are shown in Table 4. These results were consistent with the results of the two-proportional z-test. The age, race, and gender were found to be insignificant facilitators of link formation. These results are discussed in the following subsections.

Significant Facilitators

The two-proportion z-test and the Cox regression found the mutual acquaintance and the vehicle affiliation to be significant link-formation facilitators. A significant mutual acquaintance affiliation implies that two previously unconnected individuals are likely to commit a crime together if they have committed crimes with one or more shared acquaintances before. This facilitator has been well-studied in sociology (Kossinets & Watts, 2006; McPherson et al., 2001) and criminology (Coles, 2001) and studies from both domains have found that individuals tend to select new acquaintances who are “friends of friends.” In the network under study, this social mechanism also suggests that criminals operate in groups of close acquaintances and are likely to form operational cliques. According to domain experts and our previous study (Kaza et al., in press) this phenomenon is not unusual in narcotics networks, where individuals tend to have circles of trust that include friends and family members. These operational cliques enhance communication within the network and increase the capacity to act. This phenomenon is also in line with the social-closure theory (Coleman, 1990), which suggests that the greatest value is obtained from networks that are densely connected with a high level of trust among actors. This property of criminal networks is advantageous to law enforcement because it helps them form strong conspiracy cases against members of the group. Conspiracy cases generally remove more criminals from the street for a longer time as compared to individual convictions.

A significant vehicle affiliation variable implies that two criminals who have used the same vehicle for different crimes before are likely to co-offend in the future. Though this

TABLE 4. Results of the Cox regression on facilitators.

Time period	Number of nodes	p -value				
		Vehicle	Mutual acquaintance	Age	Race	Gender
2002–03	5,076	0.004*	<0.001*	0.995	0.943	0.492
2004–05	4,101	<0.001*	<0.001*	0.046*	0.212	0.514

Note. * $p < 0.05$.

affiliation has not been studied in previous dark-network studies, domain experts suggested that individuals involved in narcotics crimes often use certain vehicles. These vehicles are common to a particular gang or may have been stolen for specific purposes, like load vehicles (vehicle carrying narcotics) or scout/lookout vehicles. We included the vehicle affiliation in this study with the intuition that the affiliation may point to hidden social and operational links between two previously unrelated individuals. The statistical significance of the affiliation suggests the importance of including two-mode information in social networks. We believe that it is especially important to include such affiliation information in networks where relationships in one-mode data may be missing or incomplete. In the future, we plan to explore the use of such affiliations in the construction of social networks.

In addition to identifying significant facilitators, the Cox regression can also be used to determine the scale of influence of each of the facilitators. For example, sharing the same vehicle in different crimes increases the probability of triadic closure by a factor of 9.38, and each additional mutual acquaintance increases it by a factor of 10.79. Therefore, if two unconnected criminals have used the same vehicle in different crimes and have five mutual acquaintances then they are $9.38^1 \times 10.79^{(5-1)} \approx 127141.88$ times more likely to co-offend in the future as compared to those who do not share these affiliations. Such calculations can be used as a prediction mechanism to identify individuals who are very likely to be associated in the future based on their past activity.

Insignificant Facilitators

Many of the facilitators suggested by previous studies in this problem domain were found to be insignificant. The Cox regression analysis found homophily in age to be an insignificant factor in the link-formation process. This finding implies that individuals who are in the same age group were not necessarily likely to co-offend in future narcotics crimes. A recent study found that the criminals tend to co-offend with other individuals of a younger age (Sarnecki, 2001). Since in this study the age variable was defined based on similarity, this may have been the cause for the insignificance. Similarly, the gender affiliation may be dependent on the age of criminals. Female criminals may tend to co-offend with the same sex at a younger age and then co-offend with males as they get older (Warr, 1996). A change in the method of operationalizing the age and gender variables may lead to different results.

The race affiliation was also not found to be significant, in contrast to previous studies (Reiss & Farrington, 1991). We believe that this might be because Tucson is an immigrant city (about 1 hour from the southern border) where individuals of low socioeconomic status live in the same vicinity as immigrants of a different race. This may lead to criminal links being formed across race boundaries, and thus similarity in race may not be a good predictor of future activity. This has also been suggested by a previous study dealing with crime in an ethnically mixed environment (Sarnecki, 2001). Even though the insignificance of individual attributes

is contradictory to previous research suggesting attribute-based homophily (Louch, 2000; McPherson et al., 2001), it is possible that their effect was seen through other variables. For instance, similarity in race and age may be reflected in the mutual acquaintance affiliation. Thus, individuals having the same age group and race might have the same mutual acquaintance which, in turn, affects their link-formation likelihood.

Previous research (Milward & Raab, 2006) suggested that prison is an ideal place to develop future co-offending relationships for criminals. However, we found that inmate affiliation was insignificant for co-offending in the network studied here. This may be attributed to the inmate custody classification system implemented by Arizona Department of Corrections that attempts to separate inmates with previous affiliations (e.g., same gang membership, same criminal records) to different housing facilities. Therefore, inmates in the same housing facility that are screened by this system may have lower chance to co-offend in the future. This finding has important implications that can be used in policy decisions in other correctional facilities.

Conclusions

In this paper, we used dynamic SNA methods to examine the facilitators of link formation in a real-world social network. We studied several possible facilitators including homophily, mutual acquaintances, and various affiliations. The results showed that mutual acquaintance and shared vehicle affiliations were significant facilitators in the network under study. Homophily in age, race, and gender were not found to have a significant effect on the link-formation process in a narcotics network. We also quantified the influences of the facilitators on the triadic closure by using the hazard ratios of Cox regression analysis and used the information to calculate the likelihood of future co-offending. In addition, we examined the social causes and policy implications for the significance and insignificance of various facilitators. The set of generic dynamic SNA methods along with the corresponding statistical analyses used in our study may be applied to other types of networks to test the effect of various facilitators. This research may help the academic and practitioner community better understand the dynamics of social networks and devise effective strategies to influence their growth.

In the future, we plan to explore several directions including (a) studying evolving social networks with multiple relationships, (b) extending this dynamic analysis to other real-world social networks, and (c) adding more potential facilitators into our analysis, such as psychological and economical factors.

Acknowledgments

We would like to thank all the members of the University of Arizona Artificial Intelligence Lab for their support and assistance in this research. We also appreciate the critical

and important comments and suggestions from Detective Tim Petersen at the Tucson Police Department.

This research was supported in part by the NSF Digital Government (DG) program grant no. #9983304, NSF Information Technology Research (ITR) program grant no. #0326348, Department of Homeland Security (DHS) through the "BorderSafe" initiative grant no. #2030002, and DHS/NSF Regional Information Sharing and Collaboration grant no. #0636422.

References

- Abello, J., Pardalos, P., & Resende, M.G.C. (1999). On maximum clique problems in very large graphs. In J.M. Abello & J.S. Vitter (Eds.), *External memory algorithms* (pp. 119–130). Boston: American Mathematical Society.
- Afifi, A., Clark, V., & May, S. (2003). *Computer-aided multivariate analysis* (4th ed.). Boca Raton, FL: Chapman & Hall.
- Albert, R., & Barabasi, A.L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74, 47–97.
- Albert, R., Jeong, H., & Barabasi, A.L. (1999). Internet: Diameter of the World Wide Web. *Nature*, 401, 130–131.
- Backstrom, L., Huttenlocher, D., Kleinberg, J., & Lan, X. (2006). Group formation in large social networks: Membership, growth, and evolution. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 44–54). New York: ACM.
- Barabasi, A.L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286, 509–512.
- Barabasi, A.L., Jeong, H., Neda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311, 590–614.
- Barabasi, A.L., Jeong, H., Zéda, Z., Ravasz, E., Schubert, A., & Vicsek, T. (2002). Evolution of the social network of scientific collaborations. *Physica A*(311), 590–614.
- Bavelas, A. (1950). Communication patterns in task-oriented groups. *The Journal of the Acoustical Society of America*, 22, 725–730.
- Carley, K.M., Dombroski, M., Tsvetov, M., Reminga, J., & Kamneva, N. (2003, June). Destabilizing dynamic covert networks. Paper presented at the Eighth International Command and Control Research and Technology Symposium, Washington, DC.
- Chen, H. (2006). Intelligence and security informatics: Information systems perspective. *Decision Support Systems*, 41, 555–559.
- Coleman, J.S. (1990). *Foundations of social theory*. Cambridge, MA: Harvard University Press.
- Coles, N. (2001). It's not what you know, it's who you know that counts: Analyzing serious crime groups as social networks. *British Journal of Criminology*, 41, 580–594.
- Erdos, P., & Renyi, A. (1960). On the evolution of random graphs. *Bulletin of the International Statistical Institute*, 38, 343–347.
- Feld, S.L. (1982). Social structural determinants of similarity among associates. *American Sociological Review*, 47, 797–801.
- Freeman, C.L. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1), 35–41.
- Haythornthwaite, C. (2006). Learning and knowledge networks in interdisciplinary collaborations. *Journal of the American Society for Information Science and Technology*, 57, 1079–1092.
- Kaza, S., & Chen, H. (2009, January). Effect of inventor status on intraorganizational innovation evolution. Paper presented at the Hawaii International Conference on System Sciences (HICSS-42), Big Island, HI.
- Kaza, S., Xu, J., Marshall, B., & Chen, H. (in press). Topological analysis of criminal activity networks: Enhancing transportation security. *IEEE Transactions on Intelligent Transportation Systems*.
- Koschade, S. (2006). A social-network analysis of Jemaah Islamiyah: The applications to counterterrorism and intelligence. *Studies in Conflict & Terrorism*, 29, 589–605.
- Kossinets, G., & Watts, D.J. (2006). Empirical analysis of an evolving social network. *Science*, 311, 88–90.
- Krebs, V.E. (2001). Mapping networks of terrorist cells. *Connections*, 24(3), 43–52.
- Leenders, R.T.A.J. (1996). Evolution of friendship and best friendship choices. *Journal of Mathematical Sociology*, 21, 133–148.
- Leskovec, J., Kleinberg, J., & Faloutsos, C. (2005). Graphs over time: Densification laws, shrinking diameters, and possible explanations. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 177–187). New York: ACM.
- Lomi, A., & Pattison, P. (2006). Manufacturing relations: An empirical study of the organization of production across multiple networks. *Organization Science*, 17, 313–332.
- Louch, H. (2000). Personal network integration: Transitivity and homophily in strong-tie relations. *Social Networks*, 22, 45–64.
- McPherson, M., Smith-Lovin, L., & Cook, J.M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27, 415–444.
- Milward, H.B., & Raab, J. (2006). Dark networks as organizational problems: Elements of a theory. *International Public Management Journal*, 9, 333–360.
- Moody, J., McFarland, D., & Bender-deMoll, S. (2005). Dynamic network visualization. *American Journal of Sociology*, 110, 1206–1241.
- Nerkar, A., & Paruchuri, S. (2005). Evolution of R&D capabilities: The role of knowledge networks within a firm. *Management Science*, 51, 771–785.
- Newman, M.E.J. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences, USA*, 98, 404–409.
- Newman, M.E.J., Barabasi, A.L., & Watts, D.J. (2006). *The structure and dynamics of networks*. Princeton, NJ: Princeton University Press.
- Newman, M.E.J., Forrest, S., & Balthrop, J. (2002). E-mail networks and the spread of computer viruses. *Physical Review E*, 66(3), 035101.
- Oppenheim, A.V., & Schaffer, R.W. (1989). *Discrete-time signal processing*. Englewood Cliffs, NJ: Prentice-Hall.
- Palla, G., Barabasi, A.-L., & Vicsek, T. (2007). Quantifying social group evolution. *Nature*, 446, 664–667.
- Powell, W.W., White, D.R., Koput, K.W., & Owen-Smith, J. (2005). Network dynamics and field evolution: The growth of interorganizational collaboration in the life sciences. *American Journal of Sociology*, 110, 1132–1205.
- Raab, J., & Milward, H.B. (2003). Dark networks as problems. *Journal of Public Administration Research and Theory*, 13, 413–439.
- Rapoport, A. (1953). Spread of information through a population with socio-structural bias: II. Various models with partial transitivity. *Bulletin of Mathematical Biology*, 15, 535–546.
- Reagans, R. (2005). Preferences, identity, and competition: Predicting tie strength from demographic data. *Management Science*, 51, 1374–1383.
- Reiss, A.J. (1986). Co-offender influences on criminal careers. In A. Blumstein, J. Cohen, J. Roth, & C.A. Visher (Eds.), *Criminal Careers and "Career Criminals"* (Vol. 2, pp. 121–160). Washington, DC: National Academy Press.
- Reiss, A.J., & Farrington, D.P. (1991). Advancing knowledge about co-offending: Results from a prospective longitudinal survey of London males. *Journal of Criminal Law and Criminology*, 82, 360–395.
- Sarnacki, J. (2001). *Delinquent networks: Youth co-offending in Stockholm*. Cambridge, England: Cambridge University Press.
- Snijders, T.A.B. (1996). Stochastic actor-oriented models for network change. *Journal of Mathematical Sociology*, 21, 149–172.
- Snijders, T.A.B. (2001). The statistical evaluation of social network dynamics. In M.E. Sobel & M.P. Becker (Eds.), *Sociological methodology* (Vol. 31, pp. 361–395). London: Blackwell.
- Snijders, T.A.B. (2004). Models for longitudinal network data. In P.J. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social-network analysis* (pp. 215–246). New York: Cambridge University Press.
- Snijders, T.A.B., Steglich, C., & Schweinberger, M. (2007). Modeling the co-evolution of networks and behavior. In K. van Montfort,

- J. Aud, & A. Satora (Eds.), *Longitudinal models in the behavioral and related sciences* (pp. 41–71). Mahwah, NJ: Erlbaum.
- Sparrow, M.K. (1991). The application of network analysis to criminal intelligence: An assessment of the prospects. *Social Networks*, 13, 251–274.
- Thelwall, M. (2008). Social networks, gender, and friending: An analysis of MySpace member profiles. *Journal of the American Society for Information Science and Technology*, 59, 1321–1330.
- Warr, M. (1996). Organization and instigation in delinquent groups. *Criminology*, 34, 11–37.
- Watts, D.J., & Strogatz, S.H. (1998). Collective dynamics of “small-world” networks. *Nature*, 393, 440–442.
- Xu, J., & Chen, H. (2005). CrimeNet Explorer: A framework for criminal network knowledge discovery. *ACM Transactions on Information Systems*, 23, 201–226.
- Xu, J., Marshall, B., Kaza, S., & Chen, H. (2004). Analyzing and visualizing criminal network dynamics: A case study. In H. Chen, R. Moore, D.D. Zeng, & J. Leavitt (Eds.), *Lecture Notes in Computer Science, Vol 3073: Intelligence and Security Informatics* (pp. 359–377). Berlin: Springer.
- Yang, C.C., & Li, K.W. (2007). An associate constraint network approach to extract multilingual information for crime analysis. *Decision Support Systems*, 43, 1348–1361.