University of Kent

Kent Academic Repository

Li, Libo (2018) *Predicting Online Invitation Responses with a Competing Risk Model Using Privacy-Friendly Social Event Data.* European Journal of Operational Research, 270 (2). pp. 698-708. ISSN 0377-2217.

Downloaded from

https://kar.kent.ac.uk/70958/ The University of Kent's Academic Repository KAR

The version of record is available from

https://doi.org/10.1016/j.ejor.2018.03.036

This document version

Author's Accepted Manuscript

DOI for this version

Licence for this version

CC BY-NC-ND (Attribution-NonCommercial-NoDerivatives)

Additional information

Versions of research works

Versions of Record

If this version is the version of record, it is the same as the published version available on the publisher's web site. Cite as the published version.

Author Accepted Manuscripts

If this document is identified as the Author Accepted Manuscript it is the version after peer review but before type setting, copy editing or publisher branding. Cite as Surname, Initial. (Year) 'Title of article'. To be published in *Title of Journal* , Volume and issue numbers [peer-reviewed accepted version]. Available at: DOI or URL (Accessed: date).

Enquiries

If you have questions about this document contact ResearchSupport@kent.ac.uk. Please include the URL of the record in KAR. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from https://www.kent.ac.uk/guides/kar-the-kent-academic-repository#policies).

Predicting Online Invitation Responses with a Competing Risk Model Using Privacy-Friendly Social Event Data

Abstract Predicting people's responses to invitations is an important issue for social event management, as the decision-making process behind member responses to invitations is complicated. The purpose of this paper is to suggest a privacy-friendly method to predict whether and when people will respond to open invitations. We apply the competing risk model to predict member responses. The predictive model uses past social event participation data to infer a network structure among people who accept or reject invitations. The inferred networks collectively show the extent to which people are likely to accept or reject invitations. Validated using real datasets including 31,230 people and 8,885 events, the proposed method not only presents the variables that predict attendance (such as past attendance and social network), but also those that predict faster responses. This approach is privacy friendly, as it requires no personal information regarding people and social events (such as name, age and gender or event content). This work contributes to the predictive modeling literature as the first competing risk modeling study developed for the context of replying to a social invitation. Our findings will help event organizers predict how many people will attend events, allowing them to organize effectively.

Keywords: *decision support systems, social network analysis, survival analysis, predictive modeling*

Predicting Online Invitation Responses with a Competing Risk Model Using Privacy-Friendly Social Event Data

Libo Li *

NEOMA Business School - 59 Rue Pierre Taittinger, 51100 Reims France

* Corresponding author

Email address: libo.li@neoma-bs.fr

1. Introduction

Social event participation is an essential topic for research (Berridge, 2007). Individuals and organizations schedule many events for different purposes. People are generally expected to socialize, exchange information and expand their network during these events (Shone & Parry, 2004). Event planning is risky and time consuming, with many details to consider such as catering, staffing, and venue selection (Allen, 2005). It is almost impossible to arrange successful events without efficient, timely planning (Moyle, Kennelly, & Lamont, 2014).

The increasing use and availability of information technology and social media allows information about social events to be shared faster and easier (Weinberg & Williams, 2006). Many people develop their peer community through social media. About two billion people use social media today, with this number expected to exceed 2.4 billion in 2018 (Statista, 2015). Increasingly, people are using social media to manage their daily lives (Liu et al., 2012). Internet users organize meetings and social events via social media. Many companies have discovered market potential and also plan their events using social media.

Personal experience at social events will impact future participation, as people prefer to enrich their relationships by meeting other people with similar interests (McPherson, Smith-Lovin, & Cook, 2001). Social media are a valuable research resource, providing interesting insights into user behavior in many different domains (Borgatti, Everett, & Johnson, 2013). The main goal of this study is to predict whether and when people will respond positively to invitations. We use data from online meeting communities to predict event participation. More specifically, we use information regarding responses to past events to predict responses to new invitations.

The literature uses survey-based research to address the issue of participation prediction (Siebenaler, 2006). For example, an individual's perceived peer influence and previous experience with the event context have been found to be associated with later event participation (Siebenaler, 2006). Such studies are limited in terms of event context and sample size. Furthermore, previous studies have not fully explored the social network structures existing at these events. Despite these findings, predicting

participation is difficult for practical reasons. For example, individual attributes (perceived values) may be unavailable due to privacy concerns (Elwood & Leszczynski, 2011). Also, the context of the event may be hard to capture, or may change over time. Previous studies (Ladd, Herald-Brown, & Reiser, 2008; Siebenaler, 2006) consider that predicting participation without such information is impossible. This paper takes a more practical approach, using only past social event participation data without considering personal information or event context. We thus focuses on the past events people have attended. Since attendees tend to meet each other at different events, their relationships are likely to predict their participation at future events. People are likely to go to social events together with friends or acquaintances. Alternatively, they might decide not to attend if certain people are present. Furthermore, people may often attend the same events because of their shared interests rather than their prior relationship. We can therefore infer relationships between people based on common social events. Data concerning such relationships is valuable in predicting product adoption and customer churn (Fang, Hu, Li, & Tsai, 2013), and presumably in predicting responses to event invitations.

This paper presents an empirical study that predicts member responses to different social invitations, and constructs a social network based on past responses. This work contributes to the literature by developing a novel survival model to forecast member event participation. We validate the model using Area under the Receiver Operator Characteristic (ROC) Curve (AUC) values based on real-life datasets of 31,230 people and 8,885 events. The research model adds to the literature by using Bayesian networks to account for network variables, and a mixture model for member participation decision modeling. Thanks to these techniques, the proposed method outperforms previous methods by 24% on average, with a maximum of 52% predictive accuracy. Event organizers could use the proposed approach to classify members by participation decision and timing.

Section 2 describes the background to the study. Then, we detail the data collection procedure and research methods, before discussing the results of the study.

2. Background

2.1 Social Event Participation

Related research mainly focuses on two goals: 1) identifying factors that can help predict event participation, and 2) discovering which events people might want to attend as a result of their previous attendance. The first type of research is usually grounded in social science theories, and has found that social norms and other environmental influences can predict event participation (Ladd et al., 2008). Researchers have found that peer relations have an impact on students' activity participation at school (Ladd et al., 2008). For example, negative peer relations (i.e., peer rejection) predict lack of participation (Ladd et al., 2008). Prior experience also predicts attendances at future events (Siebenaler, 2006). The second type of research develops predictive models to recommend events. These models use recommender systems which often appear in the computer science literature. The recommender system infers user preference for events by using past event attendance data. These methods use the preferences to predict future events that users might like to attend (Purushotham & Jay Kuo, 2015). We investigate social event participation from a different angle: event organizers. Our research investigate people's responses to a particular event invitation: When will they decide to go to an event? At any given time, who is more likely to accept an invitation? How can we take into account the time it takes different people to respond to an invitation? To the best of our knowledge, no other studies have made a time-dependent prediction of user responses to an event invitation.

2.2 Social Network Analysis

People communicate in different ways. Personal connections between people form network-structured data (Leskovec, Huttenlocher, & Kleinberg, 2010). A social network consists of nodes and edges. A friendship network is formed when people identify each other as friends. People are presented as the nodes. People connect differently to all types of friends. Links between friends are the edges connecting the nodes in the network. Communication networks develop when people make phone calls and/or send emails to each other. Technical partnerships between companies form a cooperation network based on alliances and shared patents (Gilsing, Nooteboom, Vanhaverbeke, Duysters, & van

den Oord, 2008). Social network analysis (SNA) focuses on how the network structure between the objects (Fang et al., 2013) influences those within the network (Lewis, Gonzalez, & Kaufman, 2011).

When different object types are in the network, the network is considered to be heterogeneous (J. Yang, McAuley, & Leskovec, 2014). Some research articles, describe object types as "modes" (Borgatti et al., 2013). A network with two object types is a "two-mode" network. Two-mode network analysis techniques are believed to be beneficial when using two-mode data (Borgatti et al., 2013).

Network information can be quantified in many ways. For example, it is interesting to analyze information concerning people's neighbors, that is, people directly connected to them. Researchers have studied the effect of similar behavior among friends (McPherson et al., 2001). Obesity and smoking behavior tend to occur more often among friends than others. Even though direct relationships are useful for research, connectivity in the neighborhood of a node also reveals interesting findings. The (local) *clustering coefficient* (Fagiolo, 2007) quantifies the extent to which a node's neighboring nodes connect to each other. For instance, in a friendship network, the clustering coefficient is used to study how many friends know each other.

People of different social status might have distinctive patterns in their networks (Borgatti et al., 2013). Each node in the network has different connections, resulting in a complex network structure. Researchers might expect different nodes to have different roles (Borgatti et al., 2013). Some may act like a bridge, as other nodes cannot link to each other without passing through it. These nodes are more central than others. *Betweenness centrality* is a measure used to quantify how "central" a node is in the network (Gilsing et al., 2008; Marshall & Ghanekar, 2012). Such measures quantify the social network and provide input for many quantitative analyses (Baesens, Vlasselaer, & Verbeke, 2015).

Relationships in a social network can be positive or negative. Both types of relationships are important to study how people connect with each other (Leskovec et al., 2010). Other examples can be found in political science. Positive and negative relationships can be observed in political networks: "a friend's friend is a friend; an enemy's enemy is a friend" (Easley & Kleinberg, 2010; Stefaniak & Morzy, 2014). Politicians might sponsor a certain bill or vote against each other; this

allows researchers to study political alliances and antagonism (Neal, 2014). Previous voting patterns can help predict the future. Besides the binary outcome of positive and negative relationships, one can quantify the intensity of both positive and negative relationships.

2.3 Predictive Modeling Using Online Behavior Data

This study investigates the similarity of member responses to social events. The observed similarity helps predict future responses. Influential people in the network, such as celebrities, can influence others' responses.

The traditional way to collect social network data is through administered surveys. The success of social media websites, such as Facebook and Twitter, makes it possible to gather massive amounts of data on user behavior through application programming interfaces (API). These data sources help examine human behavior that would otherwise have been difficult to study.

3. Research methods

This section first describes the data collection, before describing the dependent and independent variables from the dataset used in the analysis. Next, it discusses previous methods used and finally, explains the proposed method and compares it with these previous methods.

3.1 Data collection

This study uses member responses to social event invitations on the website, Meetup (www.meetup.com). Meetup is a platform for communities to form social groups, and organize offline events. Each group holds many social events at different times. They send invitations (RSVPs) to all their members, who can reply yes or no, or can ignore the invitation. Meetup has mechanisms such as attendance approval and fees to organize member attendance.

The data collection procedure started on February 20, 2014 and covered events from February 2006 until June 2014. We collected data from social groups in three cities: New York, London, and Los Angeles. Past studies collect data by crawling the website over three months (Liu et al., 2012). Such a time frame is too short to be considered as a longitudinal setting. Longitudinal studies typically monitor a number of groups over a period of years (Ransbotham & Kane, 2011; Ren et al., 2012). It is

not feasible to cover a large number of groups over such a long period. The longitudinal studies we refer to above typically monitor between one and ten groups, making a few thousand observations. Table 1 gives an example of the invitation responses.

Table 1 An example of social event history data

EID	MID	Event opening time	Event starting time	Response	Response time
100001	1001	2014/02/17 1:00:00 PM	02/27/2014 1:00:00 PM	Yes	2014/02/20 3:05:30 PM
100001	1002	2014/02/17 1:00:00 PM	02/27/2014 1:00:00 PM	No	2014/02/20 10:30:00 AM
100001	1003	2014/02/17 1:00:00 PM	02/27/2014 1:00:00 PM	Yes	2014/02/18 5:27:45 PM
100001	1004	2014/02/17 1:00:00 PM	02/27/2014 1:00:00 PM	No response	/
100002	1001	2014/03/15 5:30:00 PM	03/22/2014 7:00:00 PM	Yes	2014/03/16 11:31:15 PM
100002	1002	2014/03/15 5:30:00 PM	03/22/2014 7:00:00 PM	Yes	2014/03/16 10:31:30 PM
100002	1003	2014/03/15 5:30:00 PM	03/22/2014 7:00:00 PM	No response	/
100002	1004	2014/03/15 5:30:00 PM	03/22/2014 7:00:00 PM	No	2014/03/16 10:37:30 PM

EID: event id
MID: member id
Event opening time: The time when an event is created
Event starting time: The time when an event starts

As indicated in Table 1, different social events are held during specific time slots. Each member's response is recorded with time stamps for each event. Based on the social event history, one can predict a member's participation. Given the social events held at time $t_1, t_2, \dots t_k, \dots t_n$, where $t_1 < t_2 < \cdots < t_k < \cdots < t_n$, we split the dataset to use events held at $t_1, t_2 \dots t_k$ for model building, and the later social events, held at $t_{k+1}, \dots t_n$, for testing the predictive results.

Table 2 Summary of the datasets

Groups	Nr of events	Nr of members	Nr of acceptances	Nr of people responded	Event time range
NyGROUP1	379	2259	6783	1582	2006/03-2014/03
NyGROUP2	638	1377	6827	988	2006/10-2014/04
NyGROUP3	284	2389	8502	1833	2006/10-2014/04
NyGROUP4	240	575	3562	449	2007/10-2014/04
NyGROUP5	294	757	6578	541	2008/03-2014/04
LaGROUP1	262	1007	5453	641	2006/11-2014/06
LaGROUP2	1614	3390	10083	1950	2006/10-2014/06
LaGROUP3	447	5251	7574	2962	2007/02-2014/06
LaGROUP4	685	324	2850	237	2007/04-2014/06
LaGROUP5	1866	1951	11878	1399	2007/06-2014/06
LdGROUP1	344	5014	12216	3501	2007/05-2014/04
LdGROUP2	415	1628	6919	928	2007/11-2014/04
LdGROUP3	298	1415	4204	906	2009/04-2014/04
LdGROUP4	897	1856	16703	1124	2009/09-2014/04
LdGROUP5	222	3043	5079	1852	2012/02-2014/04

Total number of events: 8,885
Total number of unique members: 31,230

Total number of acceptances: 115,211 Percentage of users attend 1 event: 38.0% attend between 2 and 5 events: 37.8% attend more than 5 events: 24.2%

We collected the datasets by first querying the events held in a city during the last month. For example, to obtain event data in New York, we placed a query to gather the events "city=NY" and "time window = 1 month from now" (now=February 2014, when the data was first collected). We randomly selected 5 events from the query result. Then, we further queried historical events held by these event groups. We repeated the process for Los Angeles and London. In this way, this study is longitudinal, because it gathers data over an extensive period, while covering a larger number of groups and members than traditional longitudinal community studies in management (Ransbotham & Kane, 2011; Ren et al., 2012). Table 2 summarizes the dataset for the different groups.

3.2 Overview of the variables

We used a survival model to obtain the predictions. Constructing a survival model requires both dependent and independent variables–sometimes called covariates in survival analysis (Allison, 2010). The dependent variables are *response outcome* and *timing*. The response outcome can be *yes* (=attend), *no* (=not attend) or *no reply*. *Timing* specifies when the participant responded.

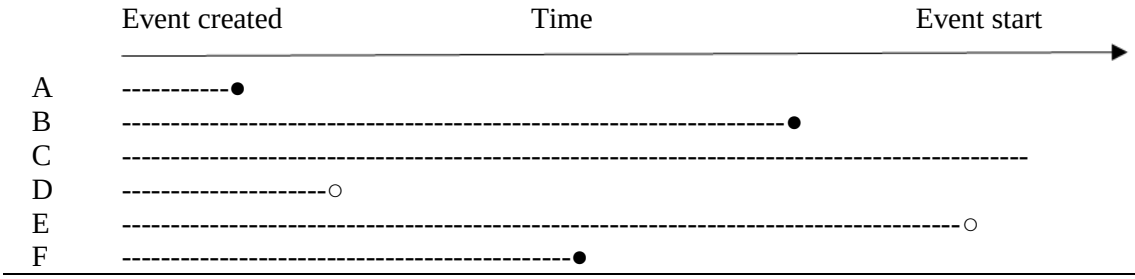

Figure 1 Time to respond

In this figure, A–F receive information about an event. They give different responses. A, B, and F, with the solid black circle, accept the invitation. D and E decline, and C does not respond. The circle indicates the response outcome, while the length of the dash line indicates the response time.

Censored data, in which the value of a measurement or observation is only partially known, is common in survival analysis. For example, people may not respond to some invitations until a certain

date. They are considered as censored data until the date they respond. In Figure 1, C never responds. C's response is censored. Censored information is an analytical problem that survival analysis deals with (Allison, 2010) since ignoring censored data might introduce bias.

We developed the independent variables from the event history. It is impossible to use the network structure directly. To predict individual responses, we extracted network information for each individual in the network. This network information is formally defined as *network features*. We also used individual *response rates*. We derived individual response rates from the number of past invitations the corresponding person accepted or declined.

The survival model uses individual characteristics such as network information to predict the response outcome. To obtain an individual's network information (sometimes known as the network features), we first determined a social network from the social event log. Then, we extracted network features for each individual in the dataset. We used these network features in the survival model (eq.9) as variables X.

For this study, we defined social groups "M_a*" and "*M_d*"* as two n-by-n adjacency matrices, where n is the total number of people who have attended past social events. In M_a, we aggregated, for each pair of people, the number of their responses that agree. In other words, in M_a, the aggregated value for person i and j is the total number of times when i and j both answered "Yes" or both answered "No". M_a reflects people's homophily. M_d represents heterophily, since it aggregates different responses. The higher the value for $M_d(i,j)$, the more node i and j disagree. In social network analysis, this is considered a projection from a two-mode network to a one-mode network. A two-mode network with social events and people is reduced into a one-mode network to help infer relationships between those people, as illustrated in Figure 2.

Building networks using event response history

1 Start with event history containing user responses, possible values = (Yes or No), total number of users = n, and n-by-n adjacency matrices M_a and M_d.

2 For i=1 to n

3 For j=1 to n

4 Agree = count the total numbers of events where user i's response = user j's response

5 M_a (i,j) = Agree

6 Disagree = count the total numbers of events where user i's response ≠ user j's response

7 M_d (i,j) = Disagree

8 End For

9 End For

Figure 2 Network construction

For each individual, we extracted network features, including *degree*, *clustering coefficient*, and *betweenness centrality*.

3.2.1 Degree

Degree (Borgatti et al., 2013) is the total number of neighbors a node has in the social network. We defined the degree of a given node i in adjacency matrices M_a and M_d in the following way (Equations 1 through 4 apply for all member i):

$$D_i(M_a) = \sum_{j=1}^{n} M_{a_{ij}} \quad \textbf{eq 1}$$

$$D_i(M_d) = \sum_{j=1}^{n} M_{d_{ij}} \quad \textbf{eq 2}$$

3.2.2 Clustering coefficient

The local *clustering coefficient* denoted as CC_i (Fagiolo, 2007) quantifies the extent to which a given node i's neighboring nodes are connected. N_i is defined as the number of edges observed between node i's neighbors. We obtained the clustering coefficient by dividing N_i by the total number of possible edges between node i's neighbors. In an undirected network, where node i has k neighbors, the total number of possible edges between node i's neighbors is $\frac{k\cdot(k-1)}{2}$.

$$CC_i = \frac{N_i}{\frac{k \cdot (k-1)}{2}} \quad \textbf{eq 3}$$

3.2.3 Betweenness centrality

Betweenness centrality (Marshall & Ghanekar, 2012), as the name suggests, shows how much a node is "in between" other nodes. The betweenness centrality Bc_i, is calculated as the number of shortest paths (geodesics) between node pair $j$and k that goes through node i, divided by the total number of shortest paths between node pair $j$and k.

$$Bc_i = \sum_{i \neq j \neq k} \frac{path(i,j,k)}{path(j,k)} \qquad \textbf{eq 4}$$

We provide a detailed discussion of these measures in the Appendix, with a numerical example for interested readers. We summarize all the variables in Table 3.

Table 3 The variables for the analysis technique

Dependent variables		*Abbreviation*
Time to respond	Continuous time variable	
Response outcome	Nominal value 1=yes 2= no 0=no response	
Independent variables		
Network features:		
Homophily degree	Numerical counts	Degree_homo
Homophily clustering coefficient	Continuous variable	Ccfs_homo
Homophily betweenness centrality	Continuous variable	Bc_homo
Heterophily degree	Numerical counts	Degree_hete
Heterophily clustering coefficient	Continuous variable	Ccfs_hete
Heterophily betweenness centrality	Continuous variable	Bc_hete
Response rates:		
Number of responses to attend events	Numerical counts	Rate_y
Number of responses not to attend events	Numerical counts	Rate_n
Total number of event invitations received	Numerical counts	Event_nr

3.3 Overview of previous approaches

Survival analysis, also known as event log history analysis, studies the occurrence of outcomes, and in this context it shows if and when a user will respond to an event invitation (Allison, 2010). Researchers in different domains might be interested in different problems. For example, whether certain medical treatments impact patient longevity; or whether an engineering process influences a

product lifetime (Allison, 2010). The survival model studies a dependent variable's change over time (Allison, 2010). Survival analysis takes into consideration the response variable and its timing (Allison, 2010).

3.3.1 Cox proportional hazard model

The survivor function (also known as the survival function) $\boldsymbol{S(t)}$ and the hazard function $\boldsymbol{h(t)}$ are two key concepts generally considered in survival analysis, and defined as follows:

$$S(t) = \Pr(T > t) \qquad \textbf{eq 5}$$

$$h(t) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \qquad \textbf{eq 6}$$

T stands for survival time, e.g. time taken to accept the invitation. The survivor function $S(t)$ is known as the probability of survival after time t, thus denoted as $\Pr(T > t)$. The hazard function $h(t)$ is the probability of accepting an invitation in the time interval $[t, t + \Delta t]$, given that the member does not accept the invitation before t. Hence, it is defined as a conditional probability distribution. The relationship between the survivor function and the hazard function can be formulated as follows:

$$S(t) = e^{-\int_0^t h(u)du} \qquad \textbf{eq 7}$$

Or alternatively

$$h(t) = -\frac{dS(t)/dt}{S(t)} \qquad \textbf{eq 8}$$

The term $\int_0^t h(u)du$ is sometimes called the cumulative hazard, as it is an "accumulation" of hazards over time. Empirically, $\int_0^t h(u)du$ can be computed using the Nelson-Aalen estimator (Borgan, 2005) as the sum of the hazards in the interval $(0, t]$. At each time point $t_i \in (0, t]$, the hazard $h(t_i)$ is the percentage of people accepting invitations at t_i, among those who have not yet accepted at t_i.

In this study, we use the Cox proportional hazard model, which is widely considered the most robust survival analysis method in different research scenarios (Allison, 2010). In Cox proportional hazard model, the hazard function $h(t)$ is based on the baseline hazard $h_0(t)$ and the covariates.

$$h(t) = h_0(t)e^{\beta X} \quad \textbf{eq 9}$$

The covariate vector *X* represents the individual's network constructs (e.g., centrality, clustering coefficient), β are the model coefficients, while $h_0(t)$ is a baseline hazard function (Cox, 1972). The baseline hazard $h_0(t)$ corresponds to the chance of an individual having 0 for all the variables (Allison, 2010). In this way, variables X are associated with the hazard function $h(t)$, and the survivor function $S(t)$.

The hazard ratio is given as e^{β}. For instance, a variable with a hazard ratio of 1.2 means that with an increase of 1 in that variable (all other variable values remaining the same), a person is 1.2 times more likely to accept the invitation. With an increase of 2, it becomes 1.2^2=1.44 times more likely. In other words, the hazard ratio indicates the "speed" with which people accept the invitation. The Cox proportional hazard model assumes the proportional hazard is constant over time, and can be extended into a time-varying model (Allison, 2010).

3.3.2 Cox model with penalization

Previous research suggests that a model using a subset of variables in a dataset sometimes outperforms the model with all variables (Aytug, 2015; Mattila & Virtanen, 2015; Miyashiro & Takano, 2015). In regression analysis, stepwise model selection is often used to obtain an optimal subset of variables. Stepwise selection has been criticized (Derksen & Keselman, 1992) for overfitting the data and producing misleading results.

In the context of survival analysis, alternative methods have been proposed to select variables. Considering the partial likelihood of the Cox proportional hazard model below:

$$L(\beta) = \prod_{s=1}^{k} \frac{e^{\beta X_s}}{\sum_{l \in R_s} e^{\beta X_l}} \quad \textbf{eq 10}$$

users can respond to the invitation at k different time points. For each of these time stamps, a fraction is calculated. The numerator is the weight of the individual responses at time s, denoted as $e^{\beta X_s}$. The denominator is the sum of the users' weights who do not respond before time s, denoted as the risk set R_s. Taking a logarithmic scale, the log likelihood function is defined as:

$$l(\beta) = \sum_{s=1}^{k} (\beta X_s - \log(\sum_{l \in R_s} e^{\beta X_l})) \quad \textbf{eq 11}$$

The coefficients are obtained using numerical optimization techniques such as the Newton-Raphson method.

New algorithms have been proposed to build models that select an optimal set of variables, such as Lasso, ridge and elastic net (Simon, Friedman, Hastie, & Tibshirani, 2011). Specifically, instead of inferring coefficients β, these new methods construct different likelihood functions:

$$\tilde{l}(\beta) = \frac{2}{n} \sum_{s=1}^{k} (\beta X_s - \log(\sum_{l \in R_s} e^{\beta X_l})) - p(\beta) \quad \textbf{eq 12}$$

The function $p(\beta)$ is known as the penalized function.

$$p(\beta) = \lambda(\alpha \sum_{i=1}^{p} |\beta_i| + \frac{1}{2}(1-\alpha) \sum_{i=1}^{p} {\beta_i}^2) \quad \textbf{eq 13}$$

The number n is the total number of observations and $\frac{2}{n}$ scales the likelihood function. The number p is the number of variables. Parameters α and λ are constants used to tune the proposed methods to adjust model coefficients for variable selection. Including the penalized function in the partial

likelihood function, the algorithm reweights the model coefficients for an optimal set of variables for prediction. Interested readers may refer to the penalized regression literature for further information (Simon et al., 2011).

Unlike their regression counterparts, survival models using penalized methods for feature selection have not yet been thoroughly investigated in a predictive modeling context (Dirick, Claeskens, & Baesens, 2015; Leow & Crook, 2016). In view of this, we include three feature selection techniques in this study: stepwise selection and elastic net implementations: "glmnet" (Friedman, Hastie, & Tibshirani, 2010), and "fastcox" (Y. Yang & Zou, 2012). The "fastcox" technique is a variation of the penalized Cox proportional hazard model, with an approximated likelihood function using the principle of majorization-minimization. Three-fold cross validation is used to select the tuning parameters for "glmnet" and "fastcox."

3.3.3 Competing risks model

Survival analysis typically concerns the transition between two states with binary outcomes; a person either responds or does not. However, in some contexts, it is useful to study multiple outcome types, which might "compete" against each other for occurrence (Kleinbaum & Klein, 2006). In this study, the response outcomes (accepting or rejecting invitations) are mutually exclusive. This leads to a specific problem in survival analysis, known as the competing risk problem.

Conventional approaches to the analysis of survival data, focus only on one type of competing risk at a time, treating other outcome types as censored. In our case, users answering "No" and censored users (who provided no answer) are treated the same. This is called the cause-specific competing risk model, where the specific causes are responses of "yes" or "no".

Fine and Gray (Fine & Gray, 1999) proposed hazard models with sub-distributions as an alternative method of analyzing competing risk data. Their approach builds separate models with respect to different outcomes. Specifically, Fine and Gray extended the standard Cox proportional hazard model with cause-specific sub-models (Fine & Gray, 1999). The likelihood function of the competing risk model extends the Cox proportional hazard model. The original risk set R_s is extended to incorporate

two populations: (1) users who have not responded by time s, and (2) users who have responded negatively by time s. While population 1 remains the same as in the original model in 3.3.1 ($w = 1\ when\ T \geq s$), population 2) includes those who respond "No" before time s. These users in the risk set R'_s are weighted by the probability of remaining unresponsive until time s.

$$L(\beta) = \prod_{s=1}^{k} \frac{e^{\beta X_s}}{\sum_{l \in R'_s} w(T, y|s) e^{\beta X_l}} \quad \textbf{eq 14}$$

$$w(T, y|s) = \begin{cases} 1, & T \geq s \\ \dfrac{G(s)}{G(T)}, & T < s \ \ AND \ \ y = NO \end{cases} \quad \textbf{eq 15}$$

G(t) is the Kaplan-Meier estimate of the survivor function of the censoring variable (Kleinbaum & Klein, 2006), and y is the response. This is the same as saying that if a user hasn't responded negatively by time s, he/she may respond "Yes" after time s ($\text{probability} = \frac{G(s)}{G(T)}$) . The population consists of "hypothetical" observations that help the survival model to assess the different user responses. This is a major difference from conventional approaches, where the competing outcomes are treated as censored. We use Fine and Gray's competing risk model in this study to take account of differences in response types.

3.3.4 Mixture Cure Models

The mixture cure model is appropriate when the population includes a sub-group that is non-susceptible to the event of interest (to accept invitations). This sub-group's survival probability is set to one, meaning that members in this sub-group will not respond at all. The mixture cure model treats data samples as a mixture of two populations: (1) those who will respond to the events and (2) those who will not (Dirick et al., 2015). In a mixture cure model the survivor function is formulated as follows:

$$s(t|X, Z) = 1 - \pi(Z) + \pi(Z) S(t|X, y = yes) \quad \textbf{eq 16}$$

Here $\pi(\cdot)$ is typically a logistic regression model predicting the probability of whether a user will respond, with a set of variables Z. The term $S(t|X, y = yes)$ is the survivor function estimating the fraction of users that will respond before time t with variables X. The mixture cure model is used to segment the studied population to improve the predictions (Dirick et al., 2015).

3.4 The proposed method - mixture cure modeling with Bayesian networks

All the survival analysis techniques introduced in the previous sections are of the "frequentist" type, where the inference relies on fixed point estimation using the maximum likelihood method. The previous "early stage prediction model" suggested using a Bayesian approach to forecast time to event problems (Fard, Wang, Chawla, & Reddy, 2016). Let us assume that event data is available until time t_c, and the study wants to forecast member response at t_f, a future time point, that is $t_f > t_c$. The following approach could be used to train a model:

$$P(y(t_c) = 1|x, t \leq t_c) = \frac{P(y(t_c) = 1, t \leq t_c) \prod_{j=1}^{p} P(x = x_j|y(t_c) = 1)}{P(x, t \leq t_c)} \qquad \textbf{eq 17}$$

The probability of observing an event in the training data $P(y(t_c) = 1|x, t \leq t_c)$ is proportional to a product of the cumulative failure function $P(y(t_c) = 1, t \leq t_c)$ and the naïve Bayes likelihood function $\prod_{j=1}^{p} P(x = x_j|y(t_c) = 1)$. The prediction at a future time point t_f can then be computed as:

$$P\big(y\big(t_f\big) = 1\big|x, t \leq t_f\big) = \frac{F(t_f) \prod_{j=1}^{p} P(x = x_j|y(t_c) = 1)}{P(x, t \leq t_f)} \qquad \textbf{eq 18}$$

It is not difficult to see that the explicit computation of $P(x, t \leq t_c)$ and $P(x, t \leq t_f)$ is unnecessary, as they stay the same for all classes, to ensure the aggregate of the probabilities is 1.

Furthermore, the naïve Bayes likelihood function described above assumes conditional independence among the variables, which might not hold for network variables. Thus, alternative methods are often needed to deliver more robust estimations (Fang et al., 2013). Tree augmented naïve Bayes (TAN) and Bayesian networks are popular alternatives (Fard et al., 2016). In the case of a TAN model:

$$P(y(t_c) = 1|x, t \leq t_c) = \frac{P(y(t_c) = 1, t \leq t_c) \prod_{j=1}^{p} P(x_j|y(t_c) = 1, x_p(j))}{P(x, t \leq t_c)} \quad \textbf{eq 19}$$

The conditional independence assumption is relaxed by computing conditional mutual information among variables. Each variable x_j is dependent on another variable $x_p(j)$. For details about TAN and Bayesian networks, please refer to (Koller & Friedman, 2009).

Previous survival model research has a number of limitations:

1. The Cox regression model likelihood function assumes a linear fixed effect relationship that is often not valid in a networked data setting
2. The "early stage prediction model" estimates the probability of an event happening in the future. Such a model lacks a component to distinguish invitees who do not respond.

To tackle this issue, we have adopted the approach in (Fard et al., 2016) by adding a logistic regression model to estimate the fraction of members not responding. Then, the survival model will learn from the remaining responding members using a Bayesian network. The result of combining eq 16, 18, and 19 is an adjusted survival function which computes the joint probability that leverages the cure fraction and posterior probability of the Bayesian networks. The adjusted survival function is summarized in eq 20, with $\tilde{t} = t_c$ for model training, and $\tilde{t} = t_f$ for model prediction.

$$s(\tilde{t}|X, Z) = 1 - \pi(Z) + \pi(Z)(1 - P(y(\tilde{t}) = 1|x, t \leq \tilde{t})) \quad \textbf{eq 20}$$

Equation 20 follows the same structure as equation 16, the only difference is that the survival proabilty in equation 16 is replaced by a Bayesian estimation used in equation 19.

The role of such a logistic regression is often known as the hurdle component in business studies (Bardhan, Oh, Zheng, & Kirksey, 2014). It precisely reflects a member's participation decision in the econometrics literature (Wooldridge, 2010). We adopt this component in this research model, which is another contribution to the predictive modeling literature on event participation forecasting. The survival probability estimates $s(\tilde{t}|X, Z)$ can be linked to indicate members' probability of accepting

invitations over time. “Survival” means “continue to not respond”, $1 - s(\tilde{t}|X, Z)$ indicates the probability of responding. We can rank the members with regard to this from the probability estimates of member attendances at a specific time. The higher a member’s probability of responding compared with other members, the higher the member’s ranking.

3.5 Experimental Setup and Evaluation Metric

We conducted experiments to perform predictive tasks using the collected datasets. We divided the datasets into training and testing sets. We used 90% of the events to build the model, and the remaining 10% as future events for prediction. In other words, among social events held at time $t_1, t_2, \ldots t_k, \ldots t_n$, where $t_1 < t_2 < \cdots < t_k < \cdots < t_n$, we chose t_k so that 90% of the events were held prior to t_k for model training, and used the other 10% events after t_k to test the model. To test the predictive accuracy, we used the AUC metric of the Receiver Operator Characteristic (ROC) curve (Fawcett, 2006). AUC is a common measure used to test time-varying predictions in survival analysis (Chen, Kodell, Cheng, & Chen, 2012; Heagerty & Zheng, 2005). To be more specific, the survival model outputs the survival score to quantify an individual’s probable response outcome (accept or not accept) at the given time. The survival score is the survival probability in most of the Cox models (as mentioned in 3.3.1), but can be time-varying probability estimates in Bayesian models (discussed in 3.4).

The AUC value is a single quantity that summarizes predictive accuracy obtained from the ROC curve (Fawcett, 2006). AUC values range from 0 to 1, while meaningful prediction ranges from 0.5 to 1; the higher the AUC, the more accurate the prediction. An AUC value of 1 means a perfect prediction.

To test time-varying predictive performance, we started from the point when the first person responds until the point when no one responds. Within this period, we built the survival model based on current responses. The survival model then predicts new responses among those who have not yet responded. We validated the prediction result every 24 hours within the period. When no new responses occurred within a 24 hour window, we did not update the prediction results within that 24 hour range. In such cases, the AUC results will be the same as the previous 24 hour period. We eliminated duplicate

results when evaluating the model performance to avoid inflated outcomes. We tested the predictive performance for each event held by each group. We tested more than 800 events and aggregated the results per group. To better understand the model performance over time, we tested the results early, mid-term and late, three stages of equal duration. Each stage consists of 1/3 of the total duration, as shown in Figure 3.

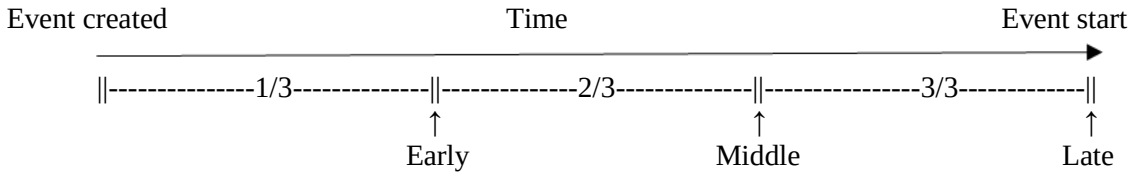

Figure 3 early, middle and late stages

Thus, we could evaluate the model not just at aggregated group level, but rather over different periods to see whether the model predictions are consistently good during each period. We included a selection of statistical models in the test, and we provide an overview of all the methods below in Table 4.

Table 4 Overview of models tested

Model specifications	Abbreviation
Logistic regression	LR
Cox proportional hazard model	coxph
Elastic-Net Regularized Cox model	glmnet
Elastic-Net Regularized Cox model with majorization-minimization	fastcox
Cox model with stepwise regression	coxSW
Proportional Subdistribution Hazards Model for Competing Risks	cmprsk
Mixture cure model	smcure
Early stage prediction	esp
Mixture cure model with Bayesian networks	bcure

4. Results

Table 5 shows the AUC values for each dataset, obtained by averaging the AUC values for predictions made at different times for the 10% testing data. The proposed **bcure** model predicts

invitation responses accurately, with an average AUC above 0.9. An AUC value higher than 0.9 is generally considered excellent (Pittman, Christensen, Caldow, Menza, & Monaco, 2007).

We used a further statistical test procedure (Benavoli, Corani, & Mangili, 2016; Demšar, 2006) to test the significance of the results. A Friedman's test (p value < 0.001) rejected the hypothesis that the predictive accuracy of all the methods are identical. We used Wilcoxon rank sum tests to compare the best performing "**bcure**" model with other candidate methods. The Wilcoxon rank sum tests rejected the hypotheses that the predictive accuracy of the candidate models were identical to that of "**bcure**" (p value < 0.001). This confirms the significance of the result: the proposed model indeed performs better than all the benchmark models.

Table 5 Summary of prediction results (AUC values)

Groups	LR	coxph	glmnet	fastcox	coxSW	cmprsk	smcure	esp	bcure
NyGROUP1	0.854	0.850	0.760	0.726	0.841	0.642	0.883	0.855	**0.997**
NyGROUP2	0.843	0.831	0.733	0.733	0.825	0.638	0.860	0.764	**0.974**
NyGROUP3	0.823	0.785	0.764	0.765	0.811	0.682	0.812	0.729	**0.983**
NyGROUP4	0.892	0.875	0.802	0.810	0.867	0.696	0.903	0.837	**0.987**
NyGROUP5	0.842	0.814	0.809	0.813	0.810	0.534	0.871	0.684	**0.957**
LaGROUP1	0.875	0.861	0.802	0.824	0.858	0.708	0.893	0.709	**0.995**
LaGROUP2	0.878	0.870	0.733	0.733	0.824	0.623	0.889	0.903	**0.992**
LaGROUP3	0.831	0.826	0.750	0.707	0.808	0.693	0.854	0.639	**0.962**
LaGROUP4	0.843	0.834	0.647	0.706	0.812	0.550	0.856	0.866	**0.987**
LaGROUP5	0.875	0.869	0.786	0.752	0.836	0.624	0.886	0.810	**0.993**
LdGROUP1	0.880	0.854	0.765	0.774	0.848	0.690	0.880	0.741	**0.976**
LdGROUP2	0.886	0.875	0.853	0.818	0.871	0.699	0.909	0.458	**0.943**
LdGROUP3	0.850	0.832	0.748	0.769	0.822	0.621	0.860	0.798	**0.989**
LdGROUP4	0.844	0.831	0.825	0.835	0.830	0.640	0.882	0.889	**0.997**
LdGROUP5	0.763	0.754	0.662	0.680	0.745	0.596	0.801	0.661	**0.989**
Average	0.852	0.837	0.763	0.763	0.827	0.642	0.869	0.756	**0.981**

The feature selection methods, such as the penalized Cox regression models and the stepwise model, provide less accurate predictions. The penalized Cox regression models, such as "glmnet" and "fastcox," rely on parameter tuning. As the number of responses changes over time, it is difficult to select optimal parameters to find the best subset of variables to predict responses. Nor does the stepwise model increase predictive accuracy. The datasets have a higher number of observations than variables (n>>p), while feature selection models generally work better when n<< p. Although feature selection could be helpful, the added value was limited in this study when using a limited number of

variables to make predictions. Additionally, high correlations among network features make feature selection difficult.

The Sub-distribution Hazards Model ("cmprsk"), which specifically includes negative responses in the computation, does not provide more accurate predictions than the cause-specific Cox model ("coxph"). In fact, members often ignore invitations if they do not want to attend. The number of negative responses is fairly small in the population, and they add very little information to increase predictive performance.

The mixture cure model ("smcure") yields the best results among Cox regression models, with an average AUC of 0.869. A Wilcoxon rank sum test (p value = 0.009) shows that the mixture cure model is significantly more accurate than the Cox proportional hazard model (Demšar, 2006).

The best performing model "bcure" uses the mixture cure model structure combined with Bayesian networks. Besides the benefit the mixture cure model brings to segment member participation decisions, the Bayesian model captures the potential dependence among the variables, against the regression counterparts. Although many different types of Bayesian network could be chosen, in this study we chose a fairly simple model, Tree augmented naïve Bayes (TAN), and already its results are impressive.

Besides the aggregated results for each group, we assessed the prediction results in early, middle and late stages.

Table 6 Summary of the prediction results from bcure over time (AUC values)

	Early	Middle	Late
NyGROUP1	0.9996	0.9983	0.9990
NyGROUP2	0.9965	0.9975	0.9985
NyGROUP3	0.9967	0.9977	0.9977
NyGROUP4	0.9952	0.9973	0.9974
NyGROUP5	0.9912	0.9971	0.9973
LaGROUP1	0.9879	0.9969	0.9968
LaGROUP2	0.9855	0.9964	0.9966
LaGROUP3	0.9871	0.9963	0.9970
LaGROUP4	0.9880	0.9966	0.9972
LaGROUP5	0.9886	0.9963	0.9960

LdGROUP1	0.9885	0.9966	0.9962
LdGROUP2	0.9855	0.9964	0.9961
LdGROUP3	0.9860	0.9965	0.9962
LdGROUP4	0.9876	0.9967	0.9962
LdGROUP5	0.9880	0.9968	0.9961
Average	0.9901	0.9969	0.9970

All the values are above 0.9. However, at the early stage the predictions were less accurate, since they are based on fewer respondents (Table 6).

We also assessed the variable importance, to show their predictive power (Grömping, 2009). Using the events from the fifteen tested groups in Table 2, we permuted each single variable. The variable importance is the average decrease in AUC value obtained by comparing the permuted model with the original model across all datasets. This method quantifies the importance of each variable in predicting responses. We assessed the relative variable importance by scaling all variable importance values sum up to 1.The higher the variable importance score, the more important a variable is in predicting responses.

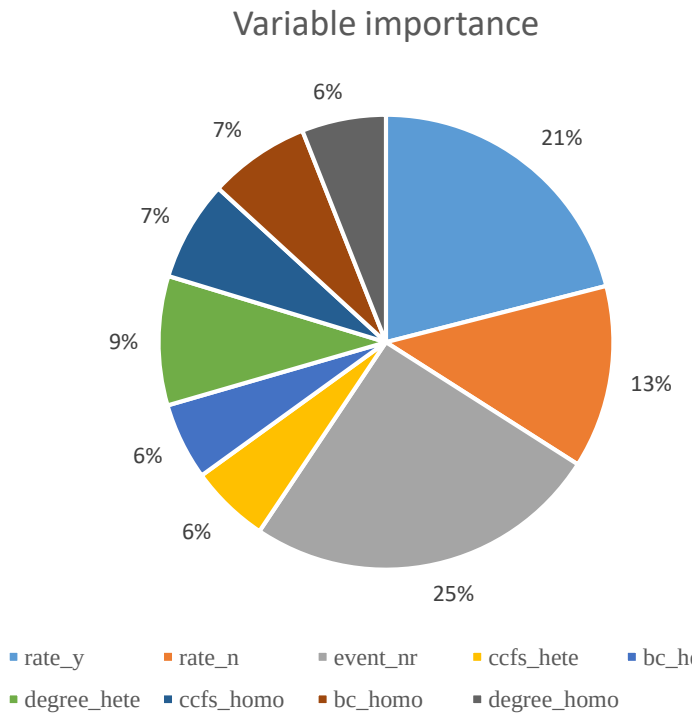

Figure 4 Variable importance scores for the proposed model

Figure 4 shows that all the variables contribute to the predictive performance. The most predictive variables are a member's previous attendance (rate_y), and the number of event invitations sent (event_nr). Positive experience of social events often suggests that members will attend next time. The longer a member remains in a social group and is exposed to more social events, the more likely he/she is to attend future events. The number of previous rejections (rate_n) also helps predict future attendance; rejection could be a sign of member unavailability. The network variables in the predictive model significantly improve the prediction results, as they contribute 40.6% of the variable importance.

5. Discussion

In this study, we propose a survival model with competing risks to predict responses to social event invitations. The paper systematically analyzed the Cox regression model and its extension, considering variable selections, competing risks, and mixture cure modeling. Based on our experiments, the mixture cure model with Bayesian networks achieves the best predictive performance.

While survival analysis has been applied to clinical studies and financial engineering, little is known of its applicability to the context of social media. The usefulness of survival analysis in new problem areas such as this need careful examination, both empirically and theoretically (Leow & Crook, 2016). We adopted the Cox model to achieve accurate predictive performance using the mixture cure modeling technique. Our findings extend the predictive modeling literature, as the proposed approach not only predicts invitation acceptance, but also when people are likely to respond. The ability to predict acceptance is crucial in many managerial areas, such as event management and scheduling. Therefore, this investigation could be useful in other areas of research, including scheduling system development (Cayirli & Veral, 2003), and general management of events (Harris, May, & Vargas, 2016).

Our study provides numerous possibilities for practitioners. Business decisions related to event resources planning could be supported by the data-driven approach proposed in this study. Identifying early respondents could facilitate event management and further assist corporate marketing campaigns

and public relationship management (Weinberg & Williams, 2006). As the proposed method relies on past responses, it could also be used to validate whether a new event still interests previous attendees and whether they are still motivated to attend. Organizers could then adapt the event settings if necessary.

If event organizers can accurately predict participation several days before an event, they can make appropriate organizational changes to improve the environment. This can contribute to a good atmosphere and attendee comfort, preventing last-minute changes. In cases of large numbers of members and limited availability, event organizers often have to schedule several events consecutively to fulfill their needs. Predicting member responses allows members to be targeted at different times for different events. This further adds value to the community, as it improves the user experience. More users could attend events to maximize community influence and to share knowledge and experience with others. An active community often contains rich information for companies to understand potential customers or enhance their product and service development (Ransbotham & Kane, 2011). The impact of these communities often extends beyond the business sector and could be useful in other fields, such as politics, ecology, and public health.

Despite the excellent predictive accuracy, this performance might be influenced by many different factors. It could be explained by other confounding factors, such as individual incidents (e.g. feeling ill, or tight deadlines at work) or specific media messages (e.g. about weather and strikes). Further research is needed to explore such phenomena, but the discussion of such confounding factors is beyond the scope of this paper.

We note that the performance of survival models is generally influenced by response times. If multiple tied response times occur in the dataset, researchers should consider those tied times in the model estimation procedure (Efron, 1977). Modern computer systems record time stamps extremely precisely, making ties extremely unlikely.

6. Limitations

This paper limits its scope by basing predictions only on online responses to social event invitations. We have ignored personal information, such as chatting behavior, group member exchanges, and

sociodemographic characteristics. While past work (Liu et al., 2012; Purushotham & Jay Kuo, 2015) has shown that individual and geographic group information generate useful insights, we ignored such information, taking a generalist approach. In practice, companies may have concerns about using more information than the invitation response, due to privacy issues. Including such information might be meaningful in particular contexts. Second, our approach does not apply to those with no past event participation data for network inference. People may join social events from other social groups and have different social event networks. We use a single social group to study the set of interested users. This comprises a limitation, as researchers can rarely access the full social network (Borgatti et al., 2013). Member participation at any given event is primarily related to the people with whom they will socialize at this event. Gathering participation information from other social groups may be difficult in real life settings (Knecht, Snijders, Baerveldt, Steglich, & Raub, 2010).

7. Conclusion

This paper proposes a competing risk model to predict social event responses using social event data and past response rates. Based on past responses, we extract network information for each individual, and use that network information to make predictions. We collected datasets from three distinctive locations to test our research model. Historical social event data on member responses proves helpful in predicting positive responses to invitations. We observe homophily and heterophily to be predictive in this study. For privacy reasons, the proposed method does not require geographical or social-demographic information. It is possible to generate knowledge about which people will respond to a social event, and when, using only historical data about other social events. This can contribute to effective event management and resource planning.

Acknowledgments

The authors thank Bart Baesens and Frank Goethals for their comments on an earlier version of this article. This research project was partially funded by IÉSEG School of Management. The authors also thank the editor-in-chief and three anonymous reviewers for their help with revising this article. This

paper also benefits from discussions with Mahtab Jahanbani Fard about her R program and related publication (Fard et al., 2016).

Appendix

Calculation of the network features

Let us denote a weighted 4 x 4 matrix M=$\begin{vmatrix} 0 & 1 & 0 & 3 \\ 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 0 \\ 3 & 2 & 0 & 0 \end{vmatrix}$, and the unweighted version of the matrix

UM =$\begin{vmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{vmatrix}$, which only considers if there is an edge (=1) or not (=0).

Degree

$$D_i(M_a) = \sum_{j=1}^{n} M_{a_{ij}} \qquad \textbf{eq A.1}$$

Since the matrix is undirected, the degree is the row/column sum of the matrix M and UM. The degree measures for $[A \quad B \quad C \quad D]$ are $[4 \quad 4 \quad 1 \quad 5]$ for the weighted version and $[2 \quad 3 \quad 1 \quad 2]$ for the unweighted version, respectively.

Local clustering coefficient

$$CC_i = \frac{N_i}{\frac{k_i \cdot (k_i - 1)}{2}} \qquad \textbf{eq A.2}$$

k_i is the number of neighbor(s) a node has. The k_i values for $[A \quad B \quad C \quad D]$ are $[2 \quad 3 \quad 1 \quad 2]$.The N_i for node i is the number of observed edges among its neighbors. Using node A as an example in the unweighted case (binary), A has two neighbors$[B \quad D]$, and there is one edge between $[B \quad D]$. Thus $N_i = 1$ and $k_i = 2$, the clustering coefficient becomes $\frac{1}{\frac{2 \cdot (2-1)}{2}} = 1$. In large graphs, the clustering coefficient will often be less than 1, since the neighbors will not establish full connections between each other. N_i is often computed in matrix algebra using the *i*th diagonal element in the third order

matrix A = M*M*M or UA = UM*UM*UM in the unweighted case. Clearly, the diagonal elements (i,i) of UM*UM are the number of neighbors, and the diagonal elements (i,i) of UM*UM*UM are the number of connections between the neighbors (N_i). Node C only has one neighbor thus the neighbor will not cluster and the clustering coefficient is set to 0. For the weighted case, the weight is normalized to the power of 1/3 before computation (Fagiolo, 2007). The M after normalization is $\begin{vmatrix} 0 & 1 & 0 & 1.44 \\ 1 & 0 & 1 & 1.26 \\ 0 & 1 & 0 & 0 \\ 1.44 & 1.26 & 0 & 0 \end{vmatrix}$, the N_i will be the (i, i) elements in $\begin{vmatrix} 0 & 1 & 0 & 1.44 \\ 1 & 0 & 1 & 1.26 \\ 0 & 1 & 0 & 0 \\ 1.44 & 1.26 & 0 & 0 \end{vmatrix}^3$. The weighted clustering coefficients for $[A \quad B \quad C \quad D]$ are respectively: $[1.81 \quad 0.61 \quad 0 \quad 1.82]$. Interested readers should refer to (Fagiolo, 2007) for a review of computing local Clustering coefficient in different disciplines, such as physics science and social behavior science.

Betweenness centrality

$$Bc_i = \sum_{i \neq j \neq k} \frac{path(i,j,k)}{path(j,k)} \qquad \textbf{eq A.3}$$

In the case of the Betweenness centrality, we see that $[A \quad B \quad D]$ connect to each other, thus none of them are "between" two others. Hence the Betweenness centrality for $[A \quad D]$ will be 0. The Betweenness centrality for node C will be 0 because there is no path going through this node. Node B will be the only node with non-zero Betweenness centrality in this case. It is in between C->D, D->C, A->C, C->A. Hence the Betweenness centrality for node B is $\frac{1}{1}+\frac{1}{1}+\frac{1}{1}+\frac{1}{1}=4$ in the unweighted case. Overall the Betweenness centrality will be $[0 \quad 4 \quad 0 \quad 0]$for nodes $[A \quad B \quad C \quad D]$. For the weighted case, the shortest path will have to take edge weights into account. In large weighted graphs, the computation of the centrality requires Dijkstra's algorithm to find the shortest path.

References

Allen, J. (2005). *Time Management for Event Planners: Expert Techniques and Time-Saving Tips for Organizing Your Workload, Prioritizing Your Day, and Taking Control of Your Schedule*. New York: Wiley.

Allison, P. D. (2010). *Survival Analysis Using SAS®: A Practical Guide, Second Edition*: SAS Institute.

Aytug, H. (2015). Feature selection for support vector machines using Generalized Benders Decomposition. *European Journal of Operational Research, 244*(1), 210-218.
Baesens, B., Vlasselaer, V. V., & Verbeke, W. (2015). *Fraud Analytics: Using Supervised, Unsupervised and Social Network Learning Techniques*: Wiley.
Bardhan, I., Oh, J.-h., Zheng, Z., & Kirksey, K. (2014). Predictive analytics for readmission of patients with congestive heart failure. *Information Systems Research, 26*(1), 19-39.
Benavoli, A., Corani, G., & Mangili, F. (2016). Should we really use post-hoc tests based on mean-ranks. *Journal of Machine Learning Research, 17*(5), 1-10.
Berridge, G. (2007). *Events design and experience*: Routledge.
Borgan, Ø. (2005). Nelson–Aalen Estimator. *Encyclopedia of Biostatistics*.
Borgatti, S. P., Everett, M. G., & Johnson, J. C. (2013). *Analyzing social networks*: SAGE Publications Limited.
Cayirli, T., & Veral, E. (2003). Outpatient scheduling in health care: a review of literature. *Production and operations management, 12*(4), 519-549.
Chen, H.-C., Kodell, R. L., Cheng, K. F., & Chen, J. J. (2012). Assessment of performance of survival prediction models for cancer prognosis. *BMC medical research methodology, 12*(1), 102.
Cox, D. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 187-220.
Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research, 7*(Jan), 1-30.
Derksen, S., & Keselman, H. (1992). Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology, 45*(2), 265-282.
Dirick, L., Claeskens, G., & Baesens, B. (2015). An Akaike information criterion for multiple event mixture cure models. *European Journal of Operational Research, 241*(2), 449-457.
Easley, D., & Kleinberg, J. (2010). Networks, Crowds, and Markets. *Cambridge University*.
Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American statistical Association, 72*(359), 557-565.
Elwood, S., & Leszczynski, A. (2011). Privacy, reconsidered: New representations, data practices, and the geoweb. *Geoforum, 42*(1), 6-15.
Fagiolo, G. (2007). Clustering in complex directed networks. *Physical Review E, 76*(2), 026107.
Fang, X., Hu, P. J.-H., Li, Z., & Tsai, W. (2013). Predicting Adoption Probabilities in Social Networks. *Information Systems Research, 24*(1), 128-145. doi: 10.1287/isre.1120.0461
Fard, M. J., Wang, P., Chawla, S., & Reddy, C. K. (2016). A bayesian perspective on early stage event prediction in longitudinal data. *IEEE Transactions on Knowledge and Data Engineering, 28*(12), 3126-3139.
Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recogn. Lett., 27*(8), 861-874. doi: 10.1016/j.patrec.2005.10.010
Fine, J. P., & Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American statistical association, 94*(446), 496-509.
Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software, 33*(1), 1.
Gilsing, V., Nooteboom, B., Vanhaverbeke, W., Duysters, G., & van den Oord, A. (2008). Network embeddedness and the exploration of novel technologies: Technological distance, betweenness centrality and density. *Research Policy, 37*(10), 1717-1731. doi: http://dx.doi.org/10.1016/j.respol.2008.08.010
Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician, 63*(4), 308-319. doi: 10.1198/tast.2009.08199
Harris, S. L., May, J. H., & Vargas, L. G. (2016). Predictive analytics model for healthcare planning and scheduling. *European Journal of Operational Research, 253*(1), 121-131. doi: http://dx.doi.org/10.1016/j.ejor.2016.02.017
Heagerty, P. J., & Zheng, Y. (2005). Survival model predictive accuracy and ROC curves. *Biometrics, 61*(1), 92-105.
Kleinbaum, D. G., & Klein, M. (2006). *Survival analysis: a self-learning text*: Springer Science & Business Media.

Knecht, A., Snijders, T. A., Baerveldt, C., Steglich, C. E., & Raub, W. (2010). Friendship and delinquency: Selection and influence processes in early adolescence. *Social Development, 19*(3), 494-514.

Koller, D., & Friedman, N. (2009). *Probabilistic graphical models: principles and techniques*.

Ladd, G. W., Herald‐Brown, S. L., & Reiser, M. (2008). Does chronic classroom peer rejection predict the development of children's classroom participation during the grade school years? *Child development, 79*(4), 1001-1015.

Leow, M., & Crook, J. (2016). The stability of survival model parameter estimates for predicting the probability of default: Empirical evidence over the credit crisis. *European Journal of Operational Research, 249*(2), 457-464. doi: https://doi.org/10.1016/j.ejor.2014.09.005

Leskovec, J., Huttenlocher, D., & Kleinberg, J. (2010). *Predicting positive and negative links in online social networks.* Paper presented at the Proceedings of the 19th international conference on World wide web.

Lewis, K., Gonzalez, M., & Kaufman, J. (2011). Social selection and peer influence in an online social network. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1109739109

Liu, X., He, Q., Tian, Y., Lee, W.-C., McPherson, J., & Han, J. (2012). *Event-based social networks: linking the online and offline social worlds.* Paper presented at the Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining.

Marshall, B., & Ghanekar, A. (2012). *Finding Betweenness in Dense Unweighted Graphs.* Paper presented at the AFIN 2012, The Fourth International Conference on Advances in Future Internet.

Mattila, V., & Virtanen, K. (2015). Ranking and selection for multiple performance measures using incomplete preference information. *European Journal of Operational Research, 242*(2), 568-579.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology, 27*, 415-444. doi: 10.2307/2678628

Miyashiro, R., & Takano, Y. (2015). Mixed integer second-order cone programming formulations for variable selection in linear regression. *European Journal of Operational Research, 247*(3), 721-731. doi: http://dx.doi.org/10.1016/j.ejor.2015.06.081

Moyle, B. D., Kennelly, M., & Lamont, M. J. (2014). Risk management and contingency planning in events: participants' reactions to the cancellation of ironman New Zealand 2012. *International Journal of Event Management Research, 9*(1), 94.

Neal, Z. (2014). The backbone of bipartite projections: Inferring relationships from co-authorship, co-sponsorship, co-attendance and other co-behaviors. *Social Networks, 39*, 84-97. doi: 10.1016/j.socnet.2014.06.001

Pittman, S., Christensen, J., Caldow, C., Menza, C., & Monaco, M. (2007). Predictive mapping of fish species richness across shallow-water seascapes in the Caribbean. *ecological modelling, 204*(1), 9-21.

Purushotham, S., & Jay Kuo, C. C. (2015). Modeling Group Dynamics for Personalized Group-Event Recommendation. In N. Agarwal, K. Xu & N. Osgood (Eds.), *Social Computing, Behavioral-Cultural Modeling, and Prediction* (Vol. 9021, pp. 405-411): Springer International Publishing.

Ransbotham, S., & Kane, G. C. (2011). Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in Wikipedia. *MIS Quarterly-Management Information Systems, 35*(3), 613.

Ren, Y., Harper, F. M., Drenner, S., Terveen, L. G., Kiesler, S. B., Riedl, J., & Kraut, R. E. (2012). Building Member Attachment in Online Communities: Applying Theories of Group Identity and Interpersonal Bonds. *Mis Quarterly, 36*(3), 841-864.

Shone, A., & Parry, B. (2004). *Successful event management: a practical handbook*: Cengage Learning EMEA.

Siebenaler, D. J. (2006). Factors that Predict Participation in Choral Music for High-School Students. *Research And Issues In Music Education, 4*(1), 1-8.

Simon, N., Friedman, J. H., Hastie, T., & Tibshirani, R. (2011). Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. *Journal of Statistical Software; Vol 1, Issue 5 (2011)*.

Statista. (2015). Facts on Social Networks Retrieved Jan, 2015, from http://www.statista.com/topics/1164/social-networks/

Stefaniak, K., & Morzy, M. (2014). Signed Graphs. In R. Alhajj & J. Rokne (Eds.), *Encyclopedia of Social Network Analysis and Mining* (pp. 1726-1734). New York, NY: Springer New York.

Weinberg, B. D., & Williams, C. B. (2006). The 2004 US Presidential campaign: Impact of hybrid offline and online 'meetup'communities. *Direct, Data and Digital Marketing Practice, 8*(1), 46-57.

Wooldridge, J. M. (2010). *Econometric analysis of cross section and panel data*.

Yang, J., McAuley, J., & Leskovec, J. (2014). *Detecting cohesive and 2-mode communities indirected and undirected networks*. Paper presented at the Proceedings of the 7th ACM international conference on Web search and data mining, New York, New York, USA.

Yang, Y., & Zou, H. (2012). A cocktail algorithm for solving the elastic net penalized Cox's regression in high dimensions. *Statistics and its Interface, 6*(2), 167-173.