Submitted to the Annals of Applied Statistics

BAYESIAN ANOMALY DETECTION METHODS FOR SOCIAL NETWORKS

By Nicholas A. Heard , David J. Weston , Kiriaki Platanioti and David J. Hand

Imperial College London

Learning the network structure of a large graph is computationally demanding, and dynamically monitoring the network over time for any changes in structure threatens to be more challenging still.

This paper presents a two-stage method for anomaly detection in dynamic graphs: the first stage uses simple, conjugate Bayesian models for discrete time counting processes to track the pairwise links of all nodes in the graph to assess normality of behavior; the second stage applies standard network inference tools on a greatly reduced subset of potentially anomalous nodes. The utility of the method is demonstrated on simulated and real data sets.

1. Introduction. Anomaly detection on graphs of social or communication networks has important security applications. The definition of a graph anomaly typically depends on the data and application of interest. Typically anomaly detection focuses on the connections amongst the graph's entities and various methods have been developed for their analysis. Examples include spectral decompositions (an area excellently summarized in von Luxburg, 2007), scan statistics (Priebe et al., 2005) and random walks (Pan et al., 2004; Tong et al., 2006). These methods are generally computationally demanding when applied to very large networks; also, in deciding upon which one to use, an explicit choice is being made on the type of anomaly sought. The interest of this paper is anomaly detection in large dynamic networks, in a context where in principle any type of anomaly should be detected. We focus on problems relating to anomalies in social networks, and present analyses of real and simulated data from this area. In each case, the network is observed over a sequence of discrete times, where each observation provides only a partial view of the full connectivity; a complete view of the network is provided by the time series as a whole.

The real data come from the European Commission Joint Research Centre's (JRC) European Media Monitor (EMM) (`http://emm.jrc.it`). EMM

*This research was supported by a DIF DTC grant.

AMS 2000 subject classifications: Primary 91D30, 90B15; secondary 62F15

Keywords and phrases: dynamic networks, Bayesian inference, counting processes, hurdle models

is a web intelligence service, providing real-time press and media summaries to Commission cabinets and services, including a breaking news and alerting service. This service requires JRC to parse each of the news documents to extract the relevant information and tag the story as belonging to a particular topic. For our analysis, JRC provided 131 weeks of Media Monitor data sourced from a collection of approved websites, starting from 1st January 2005, although this period includes a known two-week server downtime at the end of the first month. The data were extracted from news articles tagged as being related to terrorist attacks, political unrest and security. The data we receive are undirected and in a simple list format showing the date of a reported link and the names of the two individuals involved.

The simulated data come from the VAST Challenge 2008 (http://www.cs.umd.edu/hcil/VASTchallenge08); we consider the simulated cell phone data from the Mini Challenge focused in the area of social network analysis. The cell phone call records cover a fictional ten-day period on an island, narrowed to 400 unique cell phones during this period. As well as the time of each phone call and details of who phoned whom, an identifier of the cell tower from which the call originated is also given. The records should provide critical information about an important social network structure. From the results of award winning published work on this challenge (Ye et al., 2008), work which used a combination of *PageRank* (Brin and Page, 1998) and visual analytic methods, there is good reason to suspect that the major anomalous activity occurs on the eighth day and involves a list of at least eleven individuals.

2. Two-stage approach. The idea behind the method presented here is a simple one: If a social network has fundamentally changed in some important way, then in most contexts this is likely to suggest that there are some individuals who are now either communicating more or less *frequently* than usual, or communicating with *different* individuals than usual. Beyond this view there may well be much more subtle network structure to examine, but initially taking this more simple view allows good targets to be quickly identified, with the important possibility to then zoom-in and investigate such local structure.

In this paper we present a two-stage approach to dynamic anomaly detection. The first stage is a sweep of the database to identify potentially anomalous nodes in the network; in the second stage, a subgraph is constructed around this set of nodes, usually extended to include other nodes which have recently (or perhaps ever) communicated with a node in this set, and then standard network analytic tools are used to investigate structure

in this vastly reduced subnetwork.

Technically, for each pair of individuals we independently model the communications between them over time as a counting process, with the increments of the process following a Bayesian probability model. At any point in time, we test whether their relationship has changed to a degree that is statistically significant. If the derived predictive p-value falls below a fixed threshold, this represents a departure from previously modeled behavior. The node pair are then said to be anomalous and are added to the set of anomalous nodes for this period. Such an approach is statistically principled and computationally very simple. By assuming independence of the processes the method is also fully parallelizable, in the sense that each node pair is examined in isolation. This assumption of independence will be approximately acceptable only in some circumstances, and a method which seeks to relax this assumption is considered in Section 3.4.

Once a reduced subset of interesting nodes has been identified, standard network tools such as spectral clustering can be much more readily deployed; also, at this stage we are now interested in the simpler problem of characterizing structure, such as identifying clusters, rather than looking for changes in this structure, the latter being a task which requires additional metrics to be specified.

The threshold at which p-values are judged to be significant must be set by the user. In this paper we use a 0.05 threshold, but smaller or larger critical values would lead to correspondingly smaller or larger networks of potentially anomalous nodes. In practice a good threshold can be chosen to be as large as possible subject to the resulting anomaly network being of a manageable size such that follow-up investigation is feasible.

3. Discrete time counting process models. The number of communications over time are treated as simple Bayesian discrete time counting processes with conditionally independent increments. For each period in time, the number of communications between individuals will represent the current weight of their association in the network.

We first consider some different ways of counting up the communications. Then, simple Bayesian probability models are given for learning about such counting processes. Full details of these probability models and the parameterizations used are given in Supplement A.

3.1. *Pairwise, individual and total activity analysis.* For each pair of individuals (i, j), starting from time 0 when the data collection process begins let $N_{ij}(t)$ be the number of communications made from i to j up until discrete time t; alternatively, for a simpler binary view of the network, let $N_{ij}(t)$

be the number of periods in which i has communicated with j by time t. If the graph is undirected we have the simplification $N_{ij}(t) \equiv N_{ji}(t)$.

Let P_{ij} be a probability model for the increments $dN_{ij}(t) = N_{ij}(t) - N_{ij}(t-1)$ under normal circumstances. In the simplest setting, we can consider $dN_{ij}(1), dN_{ij}(2), \ldots$ as independent realisations from P_{ij}; the distribution P_{ij} corresponds to the normal mode of communication behavior for this pair of individuals. Anomalous behavior at time t, on the other hand, can be regarded as a value of $dN_{ij}(t)$ drawn from a distribution other than P_{ij}. The aim is then to detect which values of $dN_{ij}(t)$ are not draws from the unknown P_{ij}.

For a realized value of $dN_{ij}(t) = n$, we find a two-sided Bayesian p-value as the posterior probability of observing a count as extreme as n; this posterior distribution is a marginal calculation based on our revised beliefs about the unknown distribution P_{ij} in light of all other periods of data we have observed. Carefully chosen conjugate Bayesian models allow for this inferential process to be analytically tractable (see Bernardo and Smith, 1994, for details). For example, a (simplistic) parametric choice for P_{ij} could be Poisson(λ_{ij}) for unknown rate parameter $\lambda_{ij} > 0$. Completing the model specification with a gamma prior for λ_{ij} ensures the posterior predictive distribution for a future period is calculable as a simple ratio of Poisson-gamma mass functions. Where no obvious parametric form for P_{ij} exists, nonparametric Bayesian inference is available via the Dirichlet Process (Ferguson, 1973).

In the absence of specific prior information about any of the nodes, identical prior distributions are adopted for each of the node pair counting processes $N_{ij}(t)$. So in the first observation period, each node pair has the same probability of being active and hence the implied model on the whole network belongs to the well-known class of exponential random graph models (ERGMs) (Wasserman and Pattison, 1996). From the second time period onwards, however, the posterior predictive distributions will differ between node pairs according to the activity which has been observed and so here we see a departure from ERGMs.

The framework above can be regarded as an independent, *pairwise* analysis of the members of the network. If $dN_{ij}(t)$ is the *adjacency* of node i to node j at time t, then a similar *individual*-based analysis considers the

outdegree and *indegree* of node i, given by the respective increments of

$$N_{i\cdot}(t) = \sum_{j \neq i} N_{ij}(t), \tag{3.1}$$

$$N_{\cdot i}(t) = \sum_{j \neq i} N_{ji}(t). \tag{3.2}$$

These two summed processes correspond to the number of outgoing (Equation (3.1)) and incoming (Equation (3.2)) communications over time for individual i. For an undirected graph the indegree and outdegree are equivalent and Equation (3.2) is redundant. Again, we can assume exchangeable increments following similar probability models for these processes and look for outlying values in each time period.

Finally, as a highest level summary we can monitor the *degree sum* of the network over time, given by the increments of

$$N_{\cdot\cdot}(t) = \sum_{i} \sum_{j>i} N_{ij}(t) \quad \text{or} \quad N_{\cdot\cdot}(t) = \sum_{i} \sum_{j \neq i} N_{ij}(t),$$

where the two definitions correspond to undirected and directed graphs respectively. Such processes monitor the overall network activity level. Again, the same conjugate Bayesian probability models can be applied at this level.

3.2. *Parametric inference and hurdle models.* Social network graphs are typically sparse (Faloutsos et al., 2004). Particularly in larger networks, most pairs of individuals will not communicate with one another, suggesting a vanishing fraction of node pairs actually have an edge between them. This sparsity can be seen as providing an analytical advantage here, as there will be fewer non-trivial node pair relationships in the graph.

However, when the network is viewed temporally, the sparsity of the network is further increased. As we will see in the examples later, even individuals who are related will spend much of the time not communicating. This type of sparsity is problematic when modelling the counting processes, as the large number of time periods showing zero communications mean that standard exponential family distributions are inappropriate for modelling normal behavior.

We extend the exponential family probability models to their *hurdle* variants (Mullahy, 1986), which incorporate additional probability variables for determining whether or not the node pair are active in a given period t. The modelling of the process $dN_{ij}(t)$ is split into two parts, firstly a hurdle process for determining whether $dN_{ij}(t) = 0$ or $dN_{ij}(t) > 0$, and then

secondly another stochastic process governing the value taken by $dN_{ij}(t)$ at those times when the hurdle process dictates that $dN_{ij}(t) > 0$.

At time t let $A_{ij}(t)$ be the number of time periods $u \leq t$ in which $dN_{ij}(u) > 0$, meaning the node pair (i, j) were active. The increment for time t, $dA_{ij}(t) = A_{ij}(t) - A_{ij}(t-1)$ takes value 0 or 1, with $dA_{ij}(t) = 1$ indicating the pair were active in time period t. A counting process model with Bernoulli increments specifies $A_{ij}(t)$.

For times when the two individuals are active, the hurdle model also requires a second model for the increments $dN_{ij}(t) \geq 1$. We use the shifted quantities $dN_{ij}(t) - 1 \geq 0$ to define the increments of a second counting process $dB_{ij}(s)$ by the equations

$$\begin{aligned} dB_{ij}(s) &= dN_{ij}(t_s) - 1, \quad s = 1, 2, 3, \ldots, \\ t_s &= \min\{t : A_{ij}(t) = s\}, \end{aligned}$$

with resulting counting process $B_{ij}(s) = \sum_{u=1}^{s} dB_{ij}(u)$.

For the hurdle model we therefore need to specify two (typically independent) models for the counting processes $B_{ij}(\cdot)$ and $A_{ij}(\cdot)$. Assuming independence of $A_{ij}(\cdot)$ and $B_{ij}(\cdot)$, the increments of the compensator $\Lambda_{ij}(\cdot)$ for the process $N_{ij}(\cdot)$ can be expressed as

$$d\Lambda_{ij}(t) = \mathbb{E}[dN_{ij}(t)|\mathcal{H}_{t-1}] = \mathbb{E}[dA_{ij}(t)|\mathcal{H}_{t-1}](\mathbb{E}[dB_{ij}(t)|\mathcal{H}_{t-1}] + 1),$$

where for $N(t) = \{N_{ij}(t) : i \neq j\}$, $\mathcal{H}_t = \{N(u)|u = 0, 1, 2, \ldots, t\}$ is the history of the processes up until time t. Then, since

$$\begin{aligned} \mathbb{E}[dN_{ij}^2(t)|\mathcal{H}_{t-1}] &= \mathbb{E}[dA_{ij}(t)|\mathcal{H}_{t-1}](\mathbb{E}[(dB_{ij}(t) + 1)^2|\mathcal{H}_{t-1}]) \\ &= \mathbb{E}[dA_{ij}(t)|\mathcal{H}_{t-1}]\{\mathrm{Var}[dB_{ij}(t)|\mathcal{H}_{t-1}] + 1\} \\ &\quad + d\Lambda_{ij}(t)\mathbb{E}[dB_{ij}(t)|\mathcal{H}_{t-1}], \end{aligned}$$

it follows that the increments of the predictable variation of the counting process martingale $M_{ij}(t) = N_{ij}(t) - \Lambda_{ij}(t)$ satisfy

$$\begin{aligned} d\langle M_{ij}(t)\rangle &= \mathbb{E}[dA_{ij}(t)|\mathcal{H}_{t-1}]\{\mathrm{Var}[dB_{ij}(t)|\mathcal{H}_{t-1}] + 1\} \\ &\quad + d\Lambda_{ij}(t)\mathbb{E}[dB_{ij}(t)|\mathcal{H}_{t-1}] - d\Lambda_{ij}^2(t). \end{aligned}$$

These equations can be used for checking how well the models for $N_{ij}(t)$ compare in fitting the data.

3.2.1. *Bernoulli process.* The hurdle process increments $\{dA_{ij}(t)\}$ are most simply treated as a Bernoulli process

$$dA_{ij}(t) \sim \mathrm{Bernoulli}(\pi_{ij}), \quad t = 1, 2, 3, \ldots,$$

where $1 - \pi_{ij}$ can now be much greater than the zero count probability prescribed by standard exponential family models. Note that this assumes independence of the increments.

3.2.2. *Markov chain.* To enable simple dependence on the activity status in the previous time period, an alternative Markov model instead considers

$$\phi_{ij} = \Pr(dA_{ij}(t) = 1 | dA_{ij}(t-1) = 1), \tag{3.3}$$

$$\psi_{ij} = \Pr(dA_{ij}(t) = 1 | dA_{ij}(t-1) = 0). \tag{3.4}$$

For a comparable marginal probability to π_{ij} in the Bernoulli process model, note that the stationary distribution for this Markov chain implies an equilibrium probability for the pair (i, j) being active $(dA_{ij}(t) = 1)$ at any particular time t given by

$$\frac{\psi_{ij}}{1 + \psi_{ij} - \phi_{ij}}.$$

Model specification for B_{ij} can use standard exponential family distributions such as Poisson or geometric; combined with conjugate beta priors for the hurdle probabilities above, we retain fully conjugate Bayesian inference for $N_{ij}(t)$.

3.3. *Nonparametric inference.* If, even with the hurdle extensions, it is still unclear what would be a suitably simple parametric model for the number of communications, then a useful conjugate, nonparametric Bayesian alternative is the Dirichlet Process (DP) of Ferguson (1973). Using a base measure which is a small positive scalar multiple of, say, a hurdle exponential family distribution allows fully coherent but data driven inference which is largely reliant on the tail probabilities of the empirical histogram of observed counts.

3.4. *Multinomial extensions.* For directed graphs, a related approach considers using the ideas above to first model the overall counting process of activity, say, for an individual i. Then, given a particular number of communications involving individual i, we consider categorical modelling of which classes of communication they will be. The classes could correspond to other individuals in the network, or if the links are labelled with categorical types, these classes could be the link types observed.

Suppose $dN_{i\cdot}(t) = n$, so in the tth time period individual i makes n communications. Concentrating on whom the communications were with, let p_{ij} be the probability that any contact made by individual i will be to

individual j. Then assuming independence between subsequent communications, $dN_{ij}(t) \sim \text{Binomial}(n, p_{ij})$. More generally, using the vector notation $dN_{i-}(t) = (dN_{i1}(t), \ldots, dN_{i(i-1)}(t), dN_{i(i+1)}(t), \ldots)$

$$dN_{i-}(t) \sim \text{Multinomial}(n, p_{i-}).$$

Standard conjugate Bayesian inference under the multinomial model uses a Dirichlet prior for the class probabilities, see Bernardo and Smith (1994).

For a familiar goodness-of-fit hypothesis test of multinomial data, we could contrast the observed counts $dN_{ij}(t)$ with the expected np_{ij} through the familiar likelihood ratio test statistic

$$2 \sum_{j:dN_{ij}(t)>0} dN_{ij}(t) \log \left(\frac{dN_{ij}(t)}{n\mathbb{E}(p_{ij})} \right),$$

so performing a χ^2 significance test. However, such a test would not incorporate uncertainty in the overall number of communications, as it is based conditionally on observing $dN_{i\cdot}(t) = n$. Hence we obtain an augmented likelihood ratio test statistic

$$2 \left[\sum_{j:dN_{ij}(t)>0} dN_{ij}(t) \log \left(\frac{dN_{ij}(t)}{n\mathbb{E}(p_{ij})} \right) - \log\{P_{i\cdot}(dN_{i\cdot}(t) = n)\} \right]$$

which also takes into account the uncertainty in $dN_{i\cdot}(t)$.

In summary, a probability model for the overall counting process $N_{i\cdot}(t)$ for individual i, along with the multinomial model, specifies joint a distribution for the pairwise counting processes $\{N_{ij}(t)\}$. The induced dependence of these split counting processes on one another will depend on the nature of the probability model for the total number of observations $N_{i\cdot}(t)$; in the special case where this model is Poisson, the processes will be independent of one another.

4. Sequential and retrospective analyses. Typically data for a dynamic social network will arrive as an online stream. At each discrete time t we will have two inferential possibilities. The first is to decide whether the new data at t is anomalous compared to the previous data gathered, to which we give the term *sequential* analysis. For sequential analysis at time t, we are concerned with the distribution $\Pr(dN_{ij}(t)|\mathcal{H}_{t-1})$. The second possibility is to revise our decisions about all previous periods in light of the new data, to which we give the term *retrospective* analysis. For retrospective analysis at time t, we are concerned with the distributions

$\{\Pr(dN_{ij}(u)|\mathcal{H}_t/N(u)) : 1 \leq u \leq t\}$, where $\mathcal{H}_t/N(u)$ represents the history of the processes if their values at time u were not observed.

The difference between sequential and retrospective analyses is most pronounced for times near the start of the process. In sequential analysis, it is unlikely that the earliest time points will be flagged as being anomalous, since early on there are very few data points with which to compare the current observation. However, in retrospective analysis, we can look back to these early time points and now revise our opinion, in light of all that has been seen since, as to whether those periods were in fact anomalous.

Retrospective analysis can be seen as the more thorough inferential tool, as it contains sequential analysis as a special case. Sequential analysis alone is faster and more immediately relevant. Once the process has been running for sufficiently long, subsequent retrospective analyses of a much earlier time point should eventually converge in opinion, as should the retrospective and sequential analyses for more recent time points.

5. Results.

5.1. *EMM.* Here we apply our anomaly detection methods to the real EMM network data provided by JRC. The weekly counts of the contacts made by all individuals found in news website stories relating to terrorist attacks, political unrest or security between 1 January 2005 and 11 July 2007 are shown in the top panel of Figure 1.

The counting process and compensator for the activity of the whole network are shown in the second row of Figure 1. These results have been obtained from the sequential Dirichlet Process model with an uninformative negative binomial base measure (using parameter pairs (0.1,0.01) for whole network analysis, and later (0.1,0.1) for individual and pairwise analyses, see Appendix D of Supplement A); parametric analysis with a hurdle Poisson-gamma mixture with the same parameters gives very similar results. Unsurprisingly, the compensator is over-predicting activity during the known server downtime occurring in the first month, and has subsequently under-predicted the total cumulative activity for a long while after this experience. Note that the counting process martingale increments and their predictable variation (third row of Figure 1) stabilize much earlier than this, with the only major departures of the residuals from ± 2 standard deviations occurring at the corresponding spikes in the count data. These points also coincide with the lowest predictive p-values in the fourth row of Figure 1. Note that most of the remaining significant p-values (using a 0.05 threshold) in this graph correspond to highly negative martingale residuals, suggestive of further possible server downtimes.

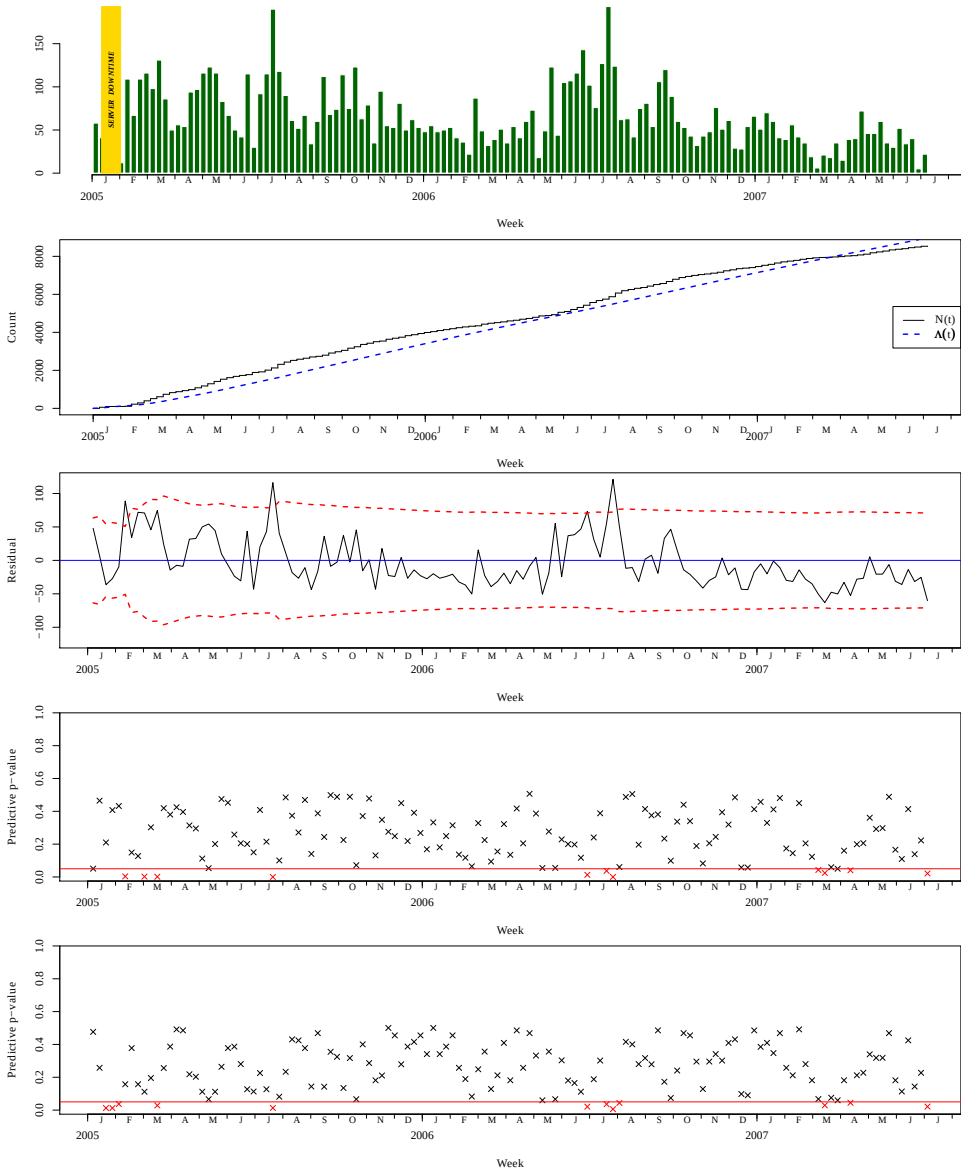

FIG 1. *Top: The number of contacts made each week by all individuals in the EMM data set. 2nd row: The counting process and compensator for the whole network activity under the sequential Dirichlet Process model. 3rd row: The martingale residuals; the dashed lines represent 2 times the square root of the predictable variation of the process. 4th row: The sequential analysis predictive p-values for the observed counts; crosses indicate values falling below a 0.05 threshold. Bottom: p-values from retrospective analysis.*

It is instructive to note that the sequential analysis p-values do not show the known server downtime to be significant. This is because we are still very much in the learning phase when the server failure occurs, and with uninformative prior beliefs the downtime is quite acceptable; rather, it is the period immediately after the downtime that is deemed anomalous. In contrast, a retrospective analysis (bottom panel of Figure 1) conducted at the end of the study correctly shows the downtime to be the anomalous period, with very small p-values.

For the pairwise and individual analyses we simply discard all data before the known server downtime. Overall there are 1,814 individuals involved in the network through the course of the observation period. The most directly connected individual is the president of the United States of America during the data collection period, George W. Bush, eventually making connections with 179 other nodes. But as mentioned earlier, social network graphs are typically sparse and here only 2,817 of the possible 1,644,391 node connections are ever made.

Across parametric and nonparametric models, the highest count of anomalous nodes identified through either individual behavior or pairwise interactions occurs on the 73rd week after the downtime, ending 28 June 2006 (see Figure 2). Interestingly, this was a time of continuing violence in the Middle East which a fortnight later would see the beginning of the the 2006 Lebanon War; it was also a week in which the Sudanese government was in disagreement with the United Nations over humanitarian involvement in the conflict in Darfur. The extended network from spectral clustering of all communicators in that week is shown in Figure 3, and important figures from the stories mentioned can be seen towards the top of the network. Nodes or links identified as anomalous, according to the individual or pairwise Poisson-gamma analyses respectively, are highlighted in red.

Taking an example pair of anomalous behaving nodes from the network in Figure 3, in Figure 4 we briefly examine the relationship of Ehud Olmert and Mahmoud Abbas. Mr Olmert became Prime Minister of Israel on 4 January 2006 (this is apparent from his growing profile in the top panel of Figure 4, and Mr Abbas has been President of the Palestinian National Authority from 15 January 2005. In the interesting week ending 28 June 2006, both individuals are showing higher than usual individual activity (although not their highest ever, see the top two panels of Figure 4); but more significantly, when viewed as a pair (bottom panel of Figure 4) they are showing a high peak of connectivity in the week ending 28 June 2006 which is unmatched at any other time over the observation period; during that week, they had met in person for the first time since Mr Olmert had taken office, and had

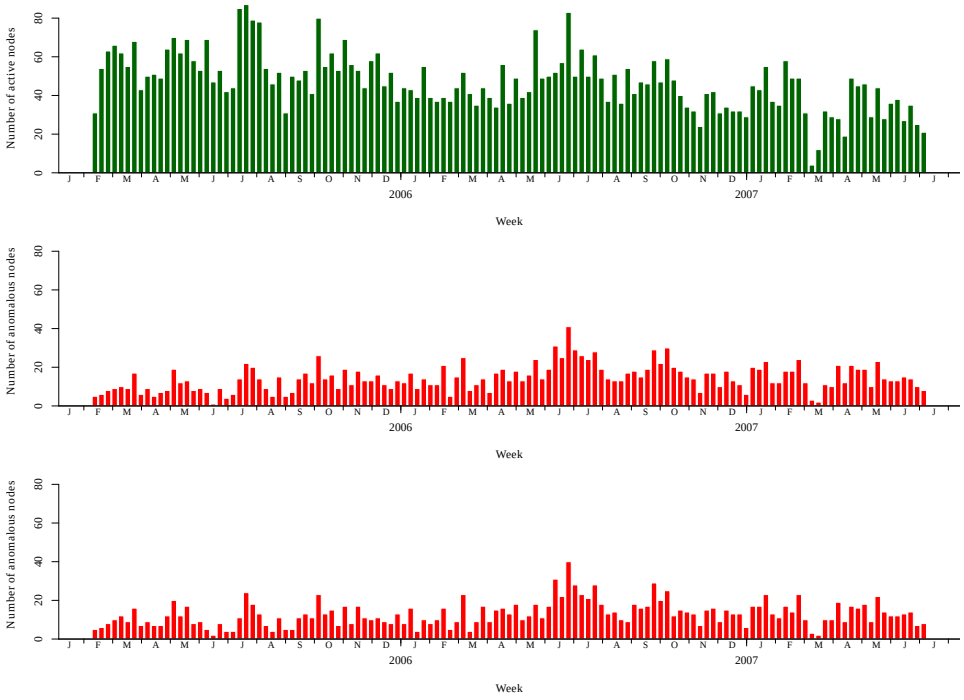

FIG 2. *The number of active nodes each week (top) and then the number of anomalous nodes found each week respectively under two models, the hurdle Poisson-gamma and the DP with Poisson-gamma base measure.*

agreed to a first official summit. Clearly this behavior should be considered anomalous, and thus they take their place in the larger network of interesting nodes in Figure 3.

FIG 3. *Network of all active individuals for the week ending 28 June 2006. Anomalous links (under sequential Dirichlet process analysis) and individuals are highlighted in red. The nodes and links of this graph are interactive in the online version of this paper.*

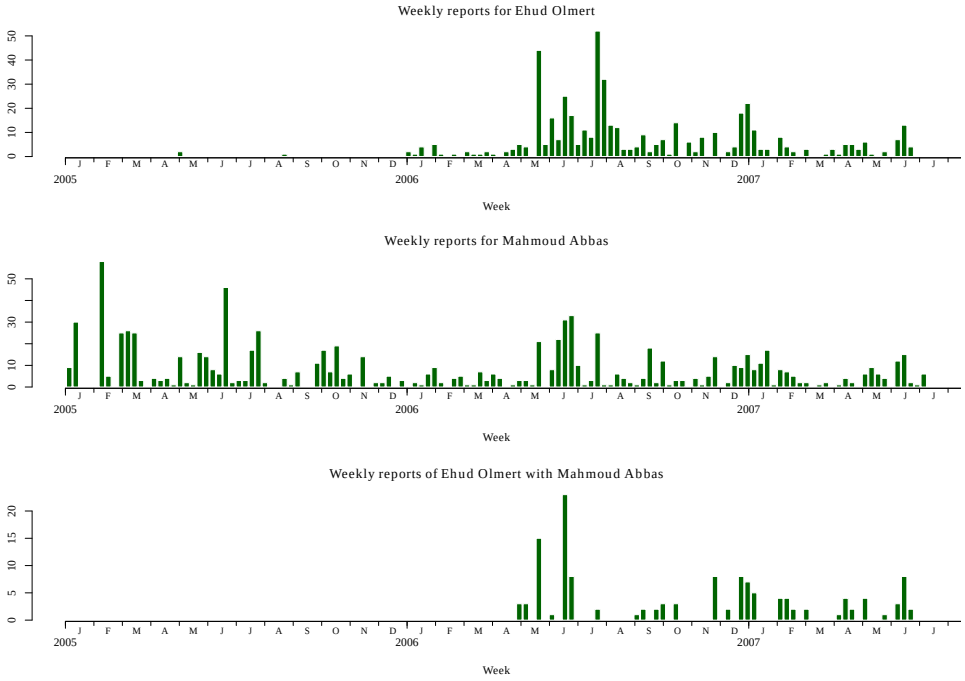

FIG 4. *Individual news report frequencies for Ehud Olmert and Mahmoud Abbas, followed by reported contacts between the pair.*

5.2. *VAST 2008.* The simulated call data from the IEEE VAST 2008 Challenge are recorded in real time, and have sufficient realism that phone calls are not seen to be made uniformly throughout the day, rather there are peak and off-peak periods. With only ten days of data, to avoid modelling of daily cyclical effects and obtain a sufficient number of homogeneous calling periods for analysis we used the histogram of daily phone calls over the ten days across the whole of the network (Figure 5, top left) to identify the broad phone call pattern; the day was then broken up into five subintervals of equal call frequencies with respect to the histogram. Thus we obtain fifty relatively homogeneous periods for our analysis (Figure 5, top right).

From the results of Ye et al. (2008) we should suspect that the major anomalous activity begins on the eighth day. This is not apparent from Figure 5, which on the bottom row also shows the number of active nodes making or receiving calls in each period and the total minutes called across the network in each period. Together, these plots do not reveal any departures from normality until the end of the ninth day. Clearly then, from a perspec-

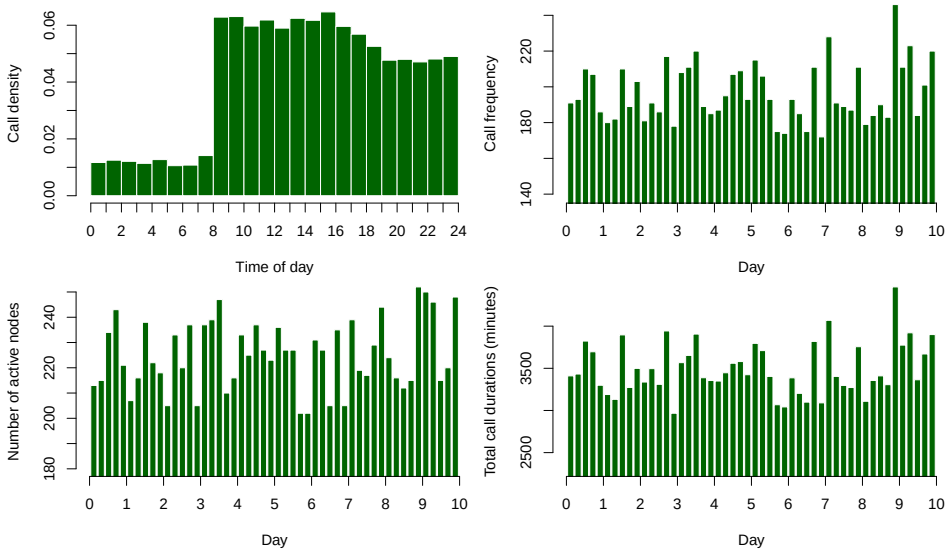

FIG 5. *Top left: Distribution of phone call start times throughout the day across the whole network for the VAST 2008 data set. Top right: Call counts across the whole network after days have been split into five homogeneous intervals. Bottom left: Number of active nodes in each subinterval. Bottom right: Total call durations across the network in each subinterval.*

tive of timely anomaly detection this anomalous behavior is not going to simply coincide with just a change in the overall activity of the network. It can be noted here that an improvement offered by our following sequential analysis over the methods used by Ye et al. (2008) is that the latter made use of the data across all of the ten days to detect the major change in the social network; our aim will be to detect the anomaly in (discretised) real time.

For these data it seems appropriate to use the Markov formulation given in Equations (3.3) and (3.4); across all individuals in the data set, the mean empirical estimates of $\phi_{i\cdot}$ and $\psi_{i\cdot}$ would be 0.63 and 0.48 respectively, so an individual making calls becomes more likely to continue making calls into the next period. Because this data set is quite short, we choose to construct empirical priors using overall means and variances across all of the call data to get broad insights into typical call volumes and their variability, which would hopefully mirror the type of prior knowledge which should be available from a domain expert. Similar but not identical results can still be achieved with uninformative priors.

Figure 6 (left) shows the number of anomalies we find in each time segment

under a sequential individual analysis of the call network nodes using a simple Bernoulli "on-off" view of node activity but incorporating the Markov assumption. There is a clear maximum at the start of the eighth day. All eight of the anomalous nodes found are from the list deduced by Ye et al. (2008) using all of the data and several combined methodologies.

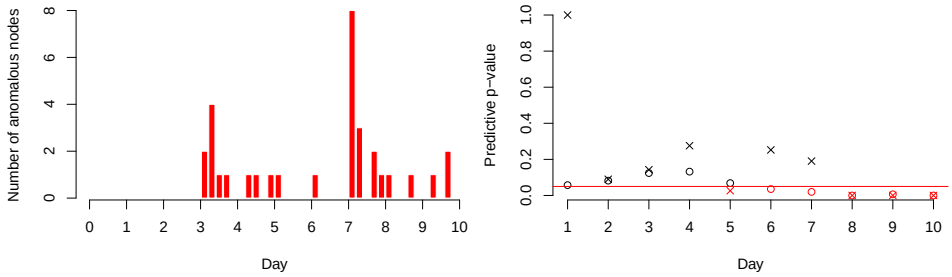

FIG 6. *Left: The number of anomalous caller IDs found in each time segment under a Markov-Bernoulli model. Right: The p-values of the anomalous ID cell-tower usage under the multinomial model; the circular points are the p-values when the number of calls is treated as random, the crosses consider the number of calls as a known quantity.*

A spectral cluster plot made using two components of the symmetric Laplacian of the historical adjacency matrix (von Luxburg, 2007) is given in Figure 7. The structure is interesting, with six of the anomalous nodes appearing together in pairs at extremes of the diagram. As shown in Ye et al. (2008), these ID pairs are each actually one individual who switches from using one cell phone to another shortly before the anomalous event occurs. Besides these three pairs, we find two of the remaining five individuals declared significant in the social network by Ye et al. (2008), although all are very much towards the center of our filtered graph and so are clearly important figures. Having found the major anomalous activity, it is a small matter to identify the remaining participants. Caller ID 200, for example, is the leader of the social network but is not detected as anomalous by our method; however, a simple investigation of the call activity of the set of anomalous nodes detected reveals ID 200 to be the most frequent communicator in the network with this group, and then the undetected ID 3 is one of only six nodes who ever communicate with the network leader; finally, as noted in Ye et al. (2008), "the person whose ID is 0 communicated with all the important people who communicated with 200".

To better understand the nature of the anomalous behavior of the circled nodes in Figure 7, we monitor their collective call activity and cell tower usage as a group using the multinomial model from Section 3.4 with

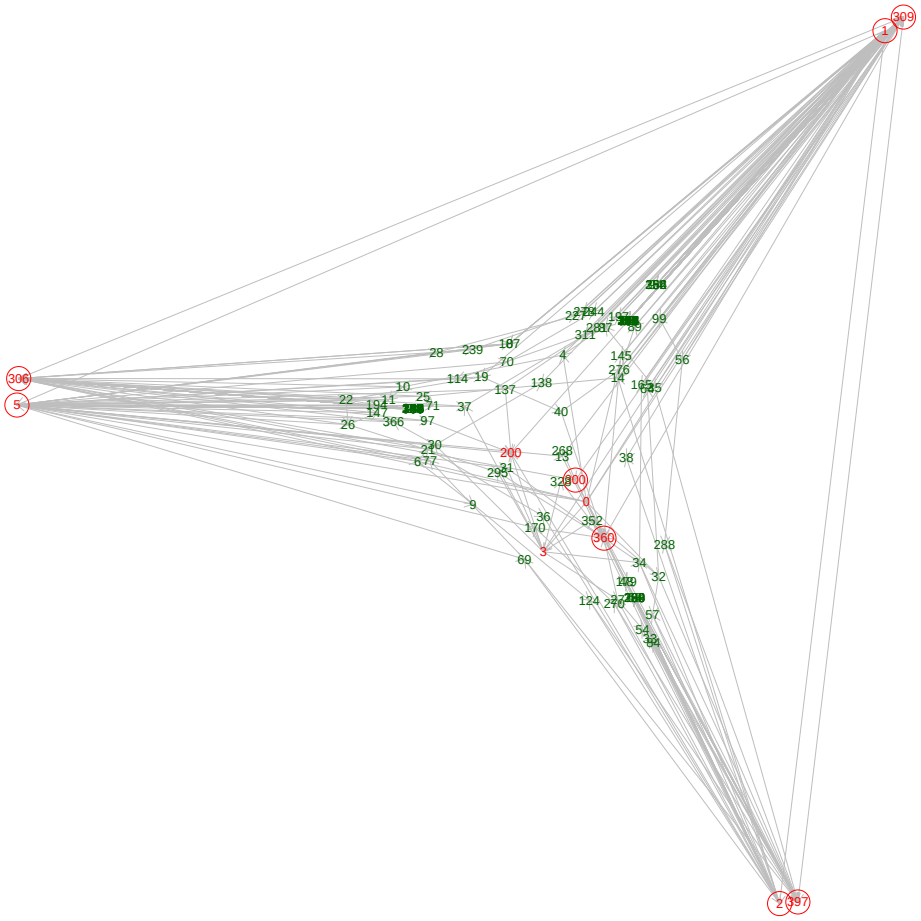

FIG 7. *Calls between cell phones during the most anomalous period, occurring at the start of the eighth day. Nodes which were identified by Ye et al. (2008) as suspicious are colored red; nodes identified as anomalous in the present analysis are circled.*

a Poisson-gamma model for the number of calls; the cell towers reveal the group's locations on the fictional island each day and so enable us to track their movements. The group's p-values from the multinomial model for each of the ten days are shown in Figure 6 (right), with the two sets of values corresponding to considering the number of calls made as either fixed or random. In obtaining these results, a Poisson-Gamma(16/9,2/9) model was used for the number of calls so that the expected number of calls equaled the number of nodes, and a flat Dirichlet $\alpha_k = 1/30$ prior used for the

multinomial model, see Supplement A for details. In terms of call volumes, the six and seventh days see a big drop in the group's call activity from an average of over eight calls per day to just one and two calls in total on those two respective days, followed by a surge of call activity on the important anomalous eighth day (30 calls) and onwards (23 and 27 calls respectively). From a cell tower perspective, day 2 is found to be fairly anomalous as the group move to using a new cell tower, number 30 (three times), whereas day 3 sees a shift in the balance of how the same set of towers are used. Besides the high call volumes, the eighth day also sees the group use seven hitherto unused cell towers - towers 7, 9, 17, 20, 21, 22 and 28, and the ninth day sees a first use of towers 2 and 12. The predictive p-values drop to near zero on these days.

6. Discussion. We have presented a simple statistical framework for monitoring dynamic social networks, by viewing the frequency of connections between node pairs as simple counting processes. Bayesian learning of the distribution of these counts enables predictive p-values to be determined for a new observation. Once a collection of interesting nodes have been identified in this way, standard network analytical methods can be used to identify the anomalous network structure.

This methodology has been successfully applied to real and simulated data sets of moderate size in this paper. Further, it has been remarked that because the methodology is mostly parallelizable and the networks are typically sparse, scaling to very large networks is feasible. For the data sets presented here the analysis is already fast; for example, for the VAST data set preprocessed into time series of length 50, identifying anomalous individual activity from the 400 IDs either as callers or through being involved in calls took 2.0 seconds. This timing is based on code run on Matlab 7.3.0 using one core on a 64-bit 1.86Ghz Xenon quad-core machine.

In both analyses, there was agreement across different count models about the peak of anomalous behavior. Spectral clustering was used to identify structure in the anomalous subnetwork, which was in agreement with knowledge from other sources.

The European Media Monitor data were from a two and a half year collection period; this could be considered as only a moderate amount of time in politics, probably overseeing at most one change in government in any represented country, for example. The simulated cell phone data covered a very short period, just ten days. The methodology presented is well suited to short or medium term modelling, as a global model is fitted across the whole time line and anomalies detected with respect to this global model.

For a longer term view, a global model is not appropriate as even normal behavior between individuals would be expected to evolve. An adaptive changepoint model with local models fitted within shorter blocks of time provides a natural extension. Such a model would lie comfortably within the Bayesian modelling paradigm, although some of the simplicity of computation enjoyed here would be lost.

Acknowledgments. This work was funded by the UK Ministry of Defence Data & Information Fusion Defence Technology Centre, project SA012 *Data mining tools for detecting anomalous clusters in network communications*. David Hand's work on this project was supported by a Royal Society Research Merit Award. We are most grateful to the EC Joint Research Centre for kindly providing us with data from their European Media Monitor.

SUPPLEMENTARY MATERIAL

Supplement A: Hurdle Exponential Family Distributions (http://stats.ma.imperial.ac.uk/~naheard/networks/supplementA.pdf). Details of the Bayesian inferential models considered in this paper.

Supplement B: Matlab/Octave code (http://stats.ma.imperial.ac.uk/~naheard/networks/supplementB.zip). Matlab code written by DJW for implementing the models used in this paper.

References.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*. Chichester: Wiley.

Brin, S. and Page, L. (1998) The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, **30**, 107–117.

Faloutsos, C., McCurley, K. S. and Tomkins, A. (2004) Connection subgraphs in social networksdetecting anomalies in graphs. In *Proceeding of SIAM International Conference on Data Mining, SIAM Workshop on Link Analysis, Counterterrorism and Security*. Newport Beach, CA, USA.

Ferguson, T. S. (1973) A Bayesian analysis of some nonparametric problems. *Annals of Statistics*, **1**, 209–230.

Mullahy, J. (1986) Specification and testing of some modified count data models. *Journal of Econometrics*, **33**, 341–365.

Pan, J.-Y., Yang, H.-J., Faloutsos, C. and Duygulu, P. (2004) Automatic multimedia cross-modal correlation discovery. In *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 653–658. New York, NY, USA: ACM.

Priebe, C. E., Conroy, J. M., Marchette, D. J. and Park, Y. (2005) Scan statistics on Enron graphs. *Computational & Mathematical Organization Theory*, **11**, 229–247.

Tong, H., Faloutsos, C. and Pan, J.-Y. (2006) Fast random walk with restart. In *ICDM'06: Sixth IEEE International Conference on Data Mining*, 613–622. Washington D.C.,USA: IEEE Computer Society.

von Luxburg, U. (2007) A tutorial on spectral clustering. *Statistics and Computing*, **17**, 395–416.

Wasserman, S. and Pattison, P. (1996) Logit models and logistic regressions for social networks: I. an introduction to markov graphs andp. *Psychometrika*, **61**, 401–425.

Ye, Q., Zhu, T., Hu, D., Wu, B., Du, N. and Wang, B. (2008) Cell Phone Mini Challenge Award: Social Network Accuracy - Exploring temporal communication in mobile call graphs. In *Proceedings of IEEE VAST Symposium*, 207–208. Piscataway, NJ: IEEE.

Department of Mathematics
Imperial College London
London SW7 2AZ
United Kingdom
E-mail: n.heard@imperial.ac.uk
E-mail: david.weston@imperial.ac.uk