Detecting Novel Discrepancies in Highly Dynamic Information Networks

James Abello[†]   Tina Eliassi-Rad[†‡]   Nishchal Devanur[†]

[†]Rutgers University   [‡]Lawrence Livermore Lab

abello@dimacs.rutgers.edu   tina@eliassi.org   ndevanur@cs.rutgers.edu

August 4, 2010

Abstract

We address the problem of detecting characteristic patterns in highly dynamic information networks (e.g., a retweet graph). We introduce a scalable approach based on set-system discrepancy. By implicitly labeling each network-edge with the sequence of times in which its two endpoints connect, we view an entire information network as a set-system. This view allows us to use combinatorial discrepancy as a mechanism to "observe" system behavior at different time scales. We illustrate our approach, called *Discrepancy-based Novelty Detector* (*DND*), on a diverse set of networks such as emails, bluetooth connections, and tweets. DND has almost linear runtime complexity in the number of pair-wise connections (i.e., communications between vertices) and linear storage complexity in the number of vertices. Examples of novel discrepancies that it detects are asynchronous connections and disagreements in the firing rates of individuals relative to the network as a whole.

Discrepancy-based Novelty Detector (DND)

On a set of vertices V, consider as input a collection of time-stamped communication pairs $\langle (x, y), t_i \rangle$ where x and y are elements of V and t_i indicates a time-stamp when the edge (x, y) was active. For each edge $e = (x, y)$, let $T_{e,t}$ denote the set of time-stamps ($t_i \leq t$) in which e is active. So, $|T_{e,t}|$ is the frequency of communications on edge $e = (x, y)$. We denote the collection of active node-pairs up to time t by $E_t = \{e = (x, y) : T_{e,t} \text{ is non-empty}\}$.

Each edge in E_t has a firing rate: $fr(e, t) = \frac{|T_{e,t}|}{t}$ and its corresponding firing sequence is $fr(e) = \langle fr(e, t) \rangle$. The firing rate of any subset E' of E_t is the sum of firing rates of the edges in E' up to time t; and its corresponding firing sequence is denoted by $fr(E') = \langle fr(E', t) \rangle$. The firing rate of a vertex x (up to a particular time t) is the sum of the firing rates of its incident edges up to that time t. So, a dynamic network is a graph sequence $\{G_t = (V, E_t)\}$ with a corresponding firing sequence $\langle fr(E_t) \rangle$. We will refer to $fr(E_t)$ as the firing rate of G_t.

Our overall approach consists of comparing the firing (and acceleration) sequence of an edge or a vertex with the firing (and acceleration) sequence of the graph in which they reside. Each G_t can be seen as the set-system: $S_t = \{T_{e,t} : T_{e,t} \text{ is a subset of a fixed ground set of time-stamps } T\}$. Therefore, a time-varying graph becomes a special set-system; and we adapt tools from set-systems' discrepancy theory[1] to study aspects of its behavior.

For the set-system $S_t = \{T_{e,t} : T_{e,t} \text{ is a subset of } \{t_0, \cdots, t\}\}$ associated with a graph G_t in the time-graph sequence $\{G_t\}$, and any two-coloring function $\chi : T \longrightarrow \{-1, 1\}$, let $\chi(T_{e,t}) = \sum \{\chi(t_i) \text{ for } t_i \in T_{e,t}\}$. This is called the *discrepancy* of the edge e at time t, with respect to the coloring χ, and abusing notation we denote it by $\chi(e, t)$. The $\chi_{discrepancy}$ of S_t is $\max\{\chi(T_{e,t}) \text{ for } T_{e,t} \in S_t\}$. The discrepancy of the set-system S_t is the minimum over all χ of $\chi_{discrepancy}(S_t)$. It follows from a fundamental result in combinatorial discrepancy that the maximum discrepancy of any of our set-systems is less than or equal to $\sqrt{2t' \ln(2m_{t'})}$, where $m_{t'}$ is the overall number of active edges up to time t'. Moreover, a random, uniform and independent coloring of $\{t_0, \cdots, t\}$ achieves this maximum. This provides us a mechanism to associate to each edge e at time t, a χ-weight in the following manner: $\chi\_wgt(e, t) = |(|\chi(e, t)| - \sqrt{2 * t * \ln(2m_t)})|$. The χ-weight of an edge up to time t measures how far is its χ value, $\chi(e, t)$, from the corresponding theoretical upper-bound on discrepancy $\sqrt{2 * t * \ln(2m_t)}$, which we refer from now on as the $sbp(t)$.

An edge e will be called χ_{novel} if its $\chi\_wgt(e, t)$ deviates substantially from the mean of the weight distribution $\chi\_wgt(E_t)$. When the discrepancy $\chi(e, t)$ is close to 0, it can be interpreted as an indication that e's "activity pattern"

---

[1]B. Chazelle. *The Discrepancy Method.* Cambridge University Press, 2001.

may be difficult to detect. On the other hand, edges with high absolute discrepancy (i.e. with low χ_wgt) exhibit an activity pattern somewhat different to the activity pattern of the entire system (again with respect to χ). Similar analysis applies to vertices and subgraphs.

Experiments

In our experiments, we used the DOJ-released email network between Enron employees from 1999 to 2002 (Enron); an MIT Reality Mining blue-tooth connections collected over 12 months (RMBT) and Twitter data with tweets from 2006 to 2009 (TWEET).

In our experimental setup, we isolate the novel edges (as defined in the previous section) when the system acceleration is at a local maxima/minima. We keep track of novel edges appearances in a Union-Find Data structure. For brevity, we mostly report on metrics over vertices which are sums of the corresponding statistics for their incident edges. All the computed statistics are functions of the edges' firing rates and accelerations.

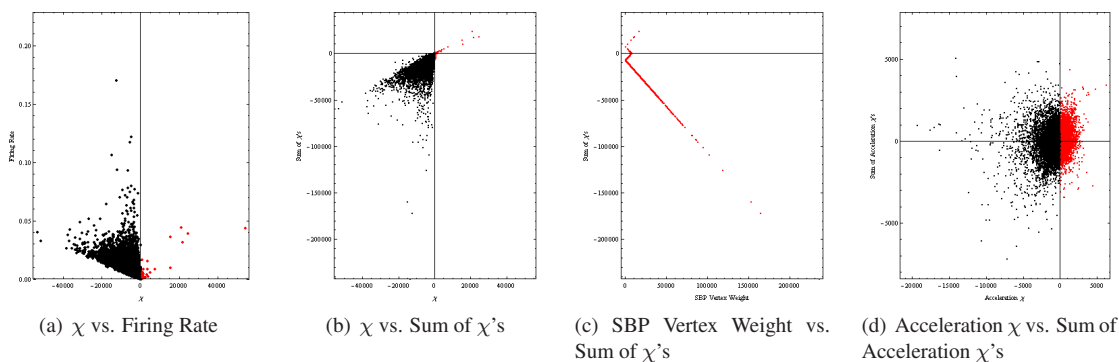| (a) χ vs. Firing Rate | (b) χ vs. Sum of χ's | (c) SBP Vertex Weight vs. Sum of χ's | (d) Acceleration χ vs. Sum of Acceleration χ's |

Figure 1: TWEET last time-stamp: scatter plots of various outputted metrics. Per plot, a red point indicates increased activity for a vertex since its last occurrence; a black point indicates the opposite (i.e., no increased activity since the vertex' last occurrence).

Figure 1 shows the scatter-plots for TWEET during the system's last time-stamp. Figure 1(a), depicting χ vs. firing rate for TWEET's last time-stamp, shows that most of the vertices (86.8%) disagree with the system (i.e., have $\chi < 0$). When observed over time, the movement of TWEET's vertices w.r.t. χ and firing rate is noticeably different than the other data sets. Initially, there are vertices that have high firing rates and χ values around zero. Then there is a phase-shift, where the firing rates for these vertices suddenly decrease and other vertices start to appear. These new vertices maintain relatively lower firing rates (because they appear late in the communication stream), however they have high negative χ values (indicating disagreement with the overall system behaviour). Figure 1(b), illustrating χ vs. $SumOf\chi$ for TWEET's last time-stamp, depicts that only 11.4% of vertices have positive values for both χ and $SumOf\chi$. This implies that only 11.4% of the vertices have behaviors that agree with the system. This result indicates that very few edges are active in each time-stamp. Figure 1(c) shows SBP vertex weight vs. $SumOf\chi$ for TWEET's last time-stamp. It illustrates a very regular pattern in the vertices, namely that their movements disagree with the system's (which is not surprising for tweets). Figure 1(d), depicting $Acceleration\chi$ vs. $SumOfAcceleration\chi$ for TWEET's last time-stamp, shows that there are no regular patterns of movement among the acceleration of vertices. Again, this is not surprising given the chaotic and heterogenous nature of communications on Twitter.

Figure 2 depicts $\chi(e,t)$ and $SBP(e,t)$ values for Enron's and RMBT's edges. The black curve corresponds to the SBP-bound on edges as a function of time. It represents the theoretical maximum set-system discrepancy. The remaining points in the scatter plot correspond to edges that at a particular time t have a particular $\chi(e,t)$ value. DND looks for those edges that reside on the narrow-band (i.e. one standard deviation around the $SBP(e,t)$, the black curve. The edges in this narrow-band correspond to "novel" edges. The same analysis holds for vertices.

Figure 3 highlights the χ vs. firing rate scatter-plots for several Enron employees who stood out in our analysis. First, we observed that a certain set of vertices show similar patterns in their movements over the same period of time. Interestingly these vertices represent people who have close working-relationship between then, such as Linda Robertson (Enron's Chief Lobbyist) and John Shelk (Enron's VP for Governmental Affairs). Another observation is
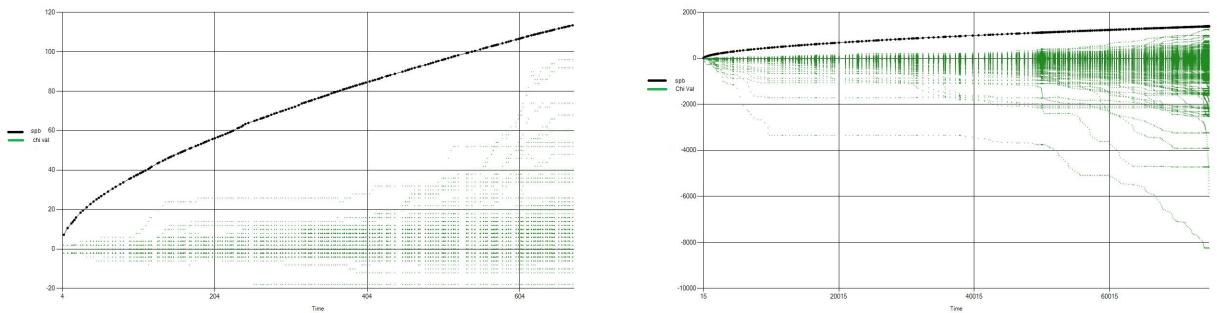
(a) Enron: $\chi(e,t)$ (in green) over time & SBP(e,t) (in black) over time   (b) RMBT: $\chi(e,t)$ (in green) over time & SBP(e,t) (in black) over time

Figure 2: Enron and RMBT: χ values on edges (green dots) and SBP values on edges (black dots) over time. The points in the narrow-band (i.e. one standard deviation) around SBP are considered "novel" based on set-system discrepancy.

that of J. Kaminski. He has a negative χ implying that his activity does not agree with that of the system. J. Kaminski was a Risk Management Expert at Enron, who warned Enron's top executives about the impending dangers.
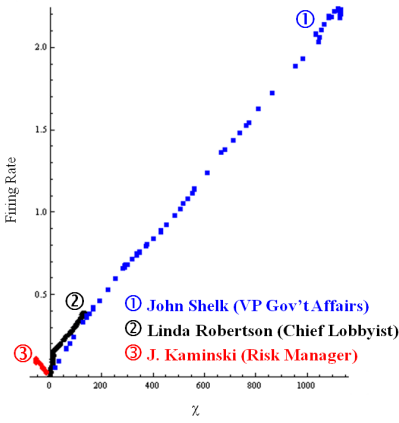

Figure 3: Enron employees who stand out based on their values for χ and firing rate. Plots #1 and #2 (in $\chi > 0$) correspond to Enron lobbyists, who show similar patterns in their emails over the same period of time. Plot #3 (in $\chi < 0$) represents J. Kaminski, a risk manager who notified Enron top executives about the potential economic bust.

Conclusions

Combinatorial set-system discrepancy is a useful mathematical construct that can isolate characteristic patterns in highly dynamic information networks. Our findings are: (1) To detect agreement in communication behavior between a graph element z and the overall network at time t, inspect $\chi(z,t)$ vs. firing rate. (2) To detect synchronicity in the system's communication at time t, examine $\chi(z,t)$ vs. $Sum of \chi(z,t)$ and $Acceleration \chi(z,t)$ vs. $Sum of Acceleration \chi(z,t)$. (3) To detect phase-shifts in communication behavior, track the discrepancy-based metrics from one time-stamp to the next. (4) To detect novel graph elements at time t, examine the graph elements whose $\chi(z,t)$ values are within few standard deviations from $SBP(z,t)$.