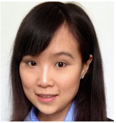
Sequential Change-Point Approach for Online Community Detection

Yao Xie
Joint work with **David Marangoni-Simonsen**

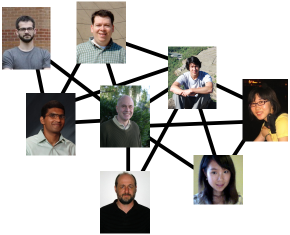
 

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

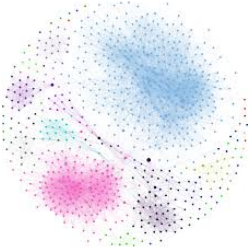
Presented at DMA Workshop, INFORMS 2014

Community

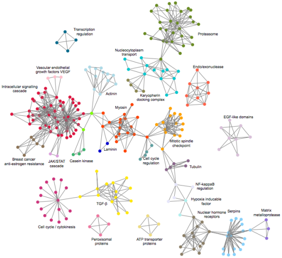
Collaboration network



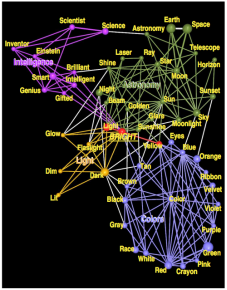
Facebook



Protein interaction network



Word association



Enron email data set


Enron Emails Reveal What a Web of Deceit Really Looks Like

ALEXIS C. MADRIGAL | JUL 13 2011, 11:54 AM ET

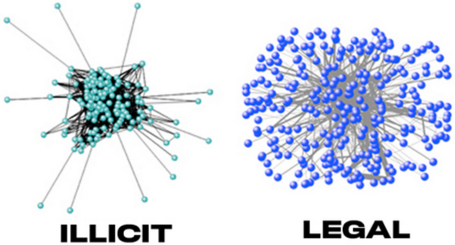
The shape that a social network takes may be a new kind of digital smoke to spot the fires of corruption within an organization

Some networks might be structurally suspicious, even if none of the content passing on it looks that way ... diagnose bad acting within a large organization. – *The Atlantic, 2011*

- 500,000 emails involving 151 unknown employees and more than 75,000 distinct addresses; each email with time stamp, sender and receiver
- between the years 1998 and 2002, record for 1,177 days

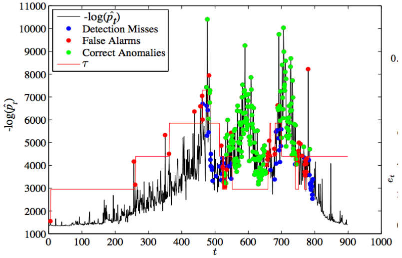
"legal project": many people are connected to many others on a project, and information is widely distributed

"illicit": information concentrated in a few hands



Emergence of a community

Starting from a certain time, anomalous email discussion topics arise between a small group of people.



Date	Significance
Dec. 1, 2000	Days before "California faces unprecedented energy alert" (Dec. 7) and energy commodity trading deregulated in Congress. (Dec. 15) [37].
May 9, 2001	"California Utility Says Prices of Gas Were Inflated" by Enron collaborator El Paso [38], blackouts affect upwards of 167,000 Enron customers [39].
Oct. 18, 2001	Enron reports \$618M third quarter loss, followed by later major correction [40].

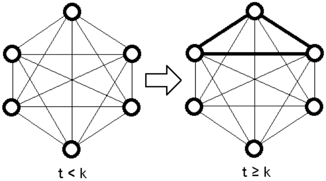
"Sequential anomaly detection in the presence of noise and limited feedback", Raginsky et al., 2013.

Online detection of community emergence

- a network with N nodes
- observe a **sequence** of independent adjacency matrices

$$X_1, X_2, \ldots$$

- $X_i \in \mathbb{R}^{N\times N}$: interaction of nodes at time i
- there may exits an unknown time s.t. after that an unknown subset of nodes interact with higher frequency



- offline version [Arias-Castro-Verzelen2014]

Sequential change-point detection approach

- H_0: X_i: Erdős-Renyi random graph

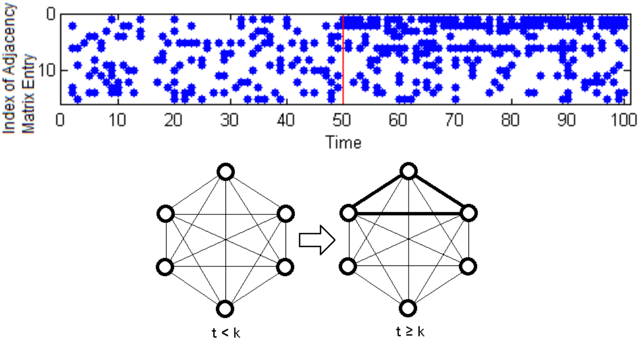
$$[X_t]_{ij} = \begin{cases} 1 & \text{w. p. } p_0 \\ 0 & \text{otherwise} \end{cases} \qquad \forall (i,j).$$

- H_1: there exists an unknown time κ such that afterwards **unknown** subset of nodes $\mathcal{S}^*$ interact more frequenctly

$$[X_t]_{ij} = \begin{cases} 1 & \text{w. p. } p_1 \\ 0 & \text{otherwise} \end{cases} \qquad \forall i,j \in \mathcal{S}^*, \quad t > \kappa,$$

$$[X_t]_{ij} = \begin{cases} 1 & \text{w. p. } p_0 \\ 0 & \text{otherwise} \end{cases} \quad \forall i \notin \mathcal{S}^* \text{ or } j \notin \mathcal{S}^*, \quad t > \kappa.$$

$$p_0 < p_1$$



- Goal: detect **emergence** of an unknown community **as quickly as possible**
- define a **stopping rule** T for sequential data such that
  - rarely raise false alarm when there is no change
  - raise alarm quickly after the change (small detection delay)

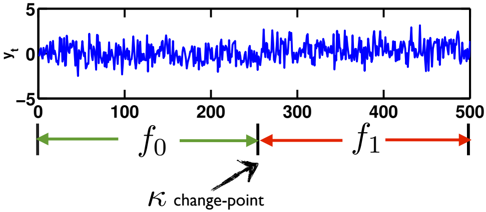
Classic change-point detection

In statistics and quality-control

- Min-max formulation: Page (54), Lorden (71)
- Bayesian: Shiryayev (63), Roberts (66)
- a sequence i.i.d. observations $y_1, y_2, \cdots \in \mathbb{R}$
- unknown change-point $\kappa > 0$.

$$
\begin{aligned}
\mathrm{H}_0 :& y_t \sim f_0, \quad t = 1, 2, \ldots \\
\mathrm{H}_1 :& y_t \sim f_0, \quad t = 1, \ldots, \kappa, \\
& y_t \sim f_1, \quad t = \kappa + 1, \ldots
\end{aligned}
$$

unknown change-point $\kappa > 0$



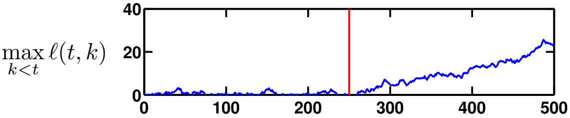
Likelihood ratio based procedure

- for a hypothesized $\kappa = k$:

$$\ell(t,k) = \log \frac{\prod_{i=1}^{k} f_0(y_i) \cdot \prod_{i=k+1}^{t} f_1(y_i)}{\prod_{i=1}^{t} f_0(y_i)} = \sum_{i=k+1}^{t} \log \frac{f_1(y_i)}{f_0(y_i)}$$

- likelihood ratio based change-point detection:

$$T = \inf\{t \geq 1 : \max_{k<t} \ell(t,k) \geq b\}$$



Normal distributions

- $f_0 = \mathcal{N}(0,1)$, $f_1 = \mathcal{N}(\mu,1)$, $\mu > 0$
- CUSUM procedure

$$T = \inf\{t : \max_{k<t} \sum_{i=k+1}^{t} (\mu y_i - \frac{\mu^2}{2}) \geq b\}$$

- when μ is **unknown**: $\hat{\mu}(k) = (\sum_{i=k+1}^{t} y_i)/(t-k)$
  GLR procedure

$$T = \inf\{t : \max_{k<t} \frac{(\sum_{i=k+1}^{t} y_i)^2}{t-k} \geq b\}$$

Likelihood ratio based statistic

- for edge (i,j), assumed change-point location $\kappa = k$, observation up to time t, likelihood ratio statistic given by

$$\ell(\kappa = k | p_1, \mathcal{S}) = \sum_{(i,j)\in\mathcal{S}} \underbrace{\sum_{m=k+1}^{t} [X_m]_{ij} \log\left(\frac{p_1}{p_0}\right) + (1 - [X_m]_{ij}) \log\left(\frac{1-p_1}{1-p_0}\right)}_{U_{k,t,p_1}^{(i,j)}}$$

- typically, we can assume p_0 **known** since it can estimated from historic data
- p_1 is usually **unknown** since it represents anomaly

Exhaustive Search (ES) method

- Approach 1: assume unknown $p_1 = \delta$
- δ: nominal value that would be important to detect

$$T_{\mathrm{ES},1} = \inf\{t : \max_{t-m_1 \le k \le t-m_0} \max_{\mathcal{S} \subset [N] : |\mathcal{S}| = s} \sum_{(i,j) \in \mathcal{S}} U_{k,t,\delta}^{(i,j)} \ge b\},$$

- exist a recursive implementation (similar to CUSUM)
- for each possible $\mathcal{S}$, calculate

$$W_{\mathcal{S},t+1} = \max\{W_{\mathcal{S},t} + \sum_{(i,j) \in \mathcal{S}} U_{t,t+1,\delta}^{(i,j)}, 0\},$$

$$T_{\mathrm{ES},1} = \inf\{t : \max_{\mathcal{S} \subset [N] : |\mathcal{S}| = s} W_{\mathcal{S},k} \ge b\}.$$

Exhaustive Search (ES) method (cont.)

- Approach 2: estimate p_1 for each hypothesize parameter values k and $\mathcal{S}$

$$\widehat{p}_1(\mathcal{S}) = \frac{2}{|\mathcal{S}|(|\mathcal{S}|-1)(t-k)} \sum_{(i,j)\in\mathcal{S}} \sum_{m=k+1}^{t} [X_m]_{ij},$$

$$T_{\mathrm{ES},2} = \inf\{t : \max_{t-m_1 \le k \le t-m_0} \max_{\mathcal{S}\subset[N]:|\mathcal{S}|=s} \sum_{(i,j)\in\mathcal{S}} U_{k,t,\widehat{p}_1(\mathcal{S})}^{(i,j)} \ge b\}.$$

- no recursive implementation
- **limitation of ES:** $\mathcal{S}$ unknown, have to search all possible subsets of $\{1,\cdots,N\}$. Number of possible subsets $|\Omega| = 2^N$, exponential in N.

Mixture method

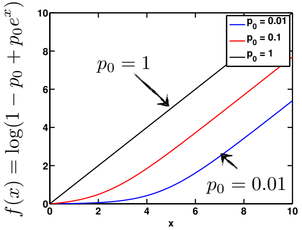
- exploit **structure**: typically community is a small subset
- assume two nodes (i, j) both in community with probability α
- α can be a guess for $|\mathcal{S}^*|/N$
- introduce indicator variable

$$Q_{ij} = \begin{cases} 1 & \text{w. p. } \alpha \\ 0 & \text{otherwise} \end{cases} \qquad \forall i, j \in \mathcal{S}^*.$$

$$\ell(\kappa = k | p_1, \mathcal{S}) = \sum_{1 \le i < j \le N} \log \Big\{ \mathbb{E}_{Q_{ij}} [(1 - Q_{ij}) + Q_{ij} \prod_{m=k+1}^{t} \frac{p_1^{[X_m]_{ij}} (1 - p_1)^{1-[X_m]_{ij}}}{p_0^{[X_m]_{ij}} (1 - p_0)^{1-[X_m]_{ij}}}] \Big\} = \sum_{1 \le i < j \le N} h(U_{k,t,p_1}^{(i,j)}).$$

Mixture method (cont.)

$$h(x) \triangleq \log\{1-\alpha+\alpha\exp(x)\}$$

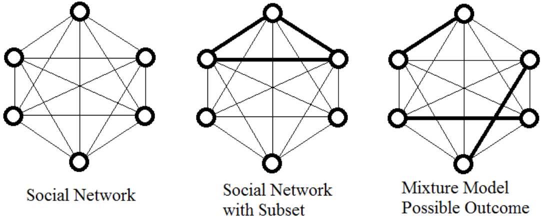


$$T_{\mathrm{Mix}} = \inf\{t : \max_{t-m_1\le k\le t-m_0} \sum_{1\le i<j\le N} h(U_{k,t,\delta}^{(i,j)}) \ge b\},$$

No search over subset $\max_{\mathcal{S}}$**.**

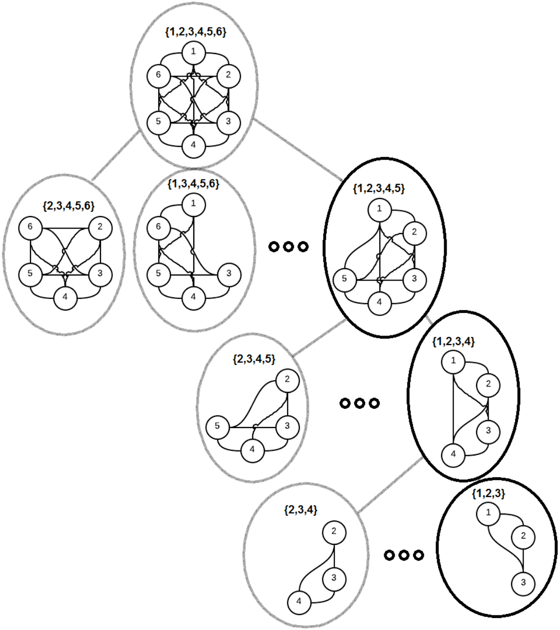
Drawback of Mixture method

- statistics of Mixture method can be gathered from "false" community
- can increase false alarm rate



- Hierarchical Mixture method (H-Mix) solves this problem by introducing dendrogram decomposition of the graph

Hierarchical Mixture method (H-Mix)



Algorithm 1 Hierarchical Mixture Method

```
Input: {X_m}_{m=1}^t, X_m ∈ R^{N×N}
Output: {P_k}_{k=1}^t ∈ R^t, a set of statistics for each hypothesized changepoint location k.
for k = 1 → t do
    S = [[N]]
    while |S| > s do
        i* = argmax_{i∈S} M (S\{i})
        S = S\{i*}
    end while
    P_k = M(S)
end for
```

Complexity

Table : Complexities of algorithms under various conditions regarding k and N.

	$\lvert\mathcal{S}\rvert \gg N/2$	$\lvert\mathcal{S}\rvert \ll N/2$	$\lvert\mathcal{S}\rvert \sim N/2$
Exhaustive Search	$\mathcal{O}(N^{N-\lvert\mathcal{S}\rvert})$	$\mathcal{O}(N^{\lvert\mathcal{S}\rvert})$	$\mathcal{O}(2^{\frac{\lvert\mathcal{S}\rvert}{2}})$
Mixture Model	$\mathcal{O}(N^2)$	$\mathcal{O}(N^2)$	$\mathcal{O}(N^2)$
Hierarchical Mixture	$\mathcal{O}(N^3)$	$\mathcal{O}(N^4)$	$\mathcal{O}(N^4)$

Choice of b

Choice of threshold b involves a tradeoff between **ARL** and **EDD**:

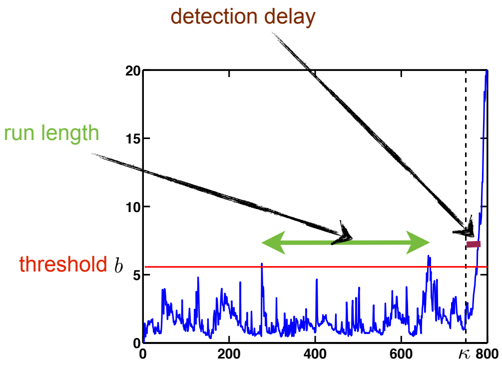
ARL (average run length) (captures false-alarm-rate)

- usually choose b to make ARL large $\sim$ 5000, 10000
- for large N simulating ARL via Monte Carlo is hard
- accurate theoretical approximation for ARL is highly valuable

EDD (expected detection delay)

- a relatively small number ~ 10
- theoretical approximation provides useful insight

Performance metrics



- average run length (ARL):

$$\mathbb{E}^{\infty}\{T\}$$

- expected detection delay (EDD):

$$\sup_k \text{ess sup } \mathbb{E}^k\{T - k | T > k\}$$

Theoretical results

We obtain analytical expression for ARL of Mixture method

Theorem

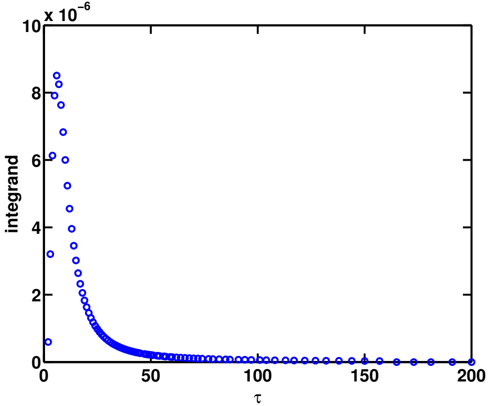
When $b \to \infty$, an upper approximation to the ARL $\mathbb{E}^{\infty}[T_{\mathrm{mix}}]$ of the Mixture method with known p_1 is given by:

$$ARL_{\mathrm{UA}} = \left[\int_{\sqrt{2N/m_1}}^{\sqrt{2N/m_0}} \frac{y\nu^2(y\sqrt{\gamma(\theta_y)})}{H(N,\theta_y)} dy \right]^{-1}, \qquad (1)$$

and a lower approximation to the ARL is given by:

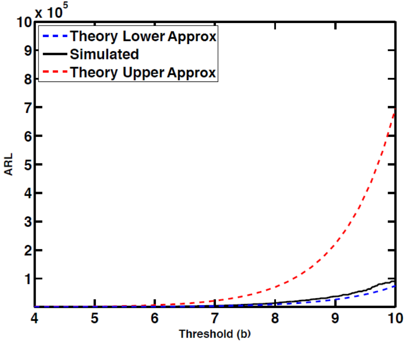
$$ARL_{\mathrm{LA}} = \left[\sum_{\tau=m_0}^{m_1} \frac{2N\nu^2(2N\sqrt{\gamma(\theta_\tau)}/\tau^2)}{\tau^2 H(N,\theta_\tau)} \right]^{-1}, \qquad (2)$$

- expressions can be evaluated explicitly
- no Monte Carlo simulation needed



only a few τ values play role in the summation

- accurate approximation
- can be used to determine threshold b for given ARL



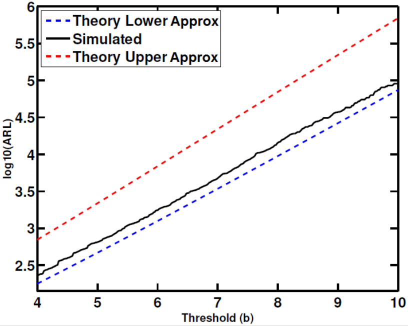
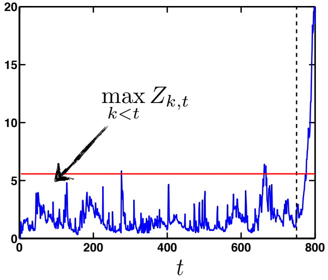


Table : Theoretical vs. simulated thresholds for $p_0 = 0.3$, $p_1 = 0.8$, and $N = 6$. The threshold b calculated using theory is very close to the corresponding threshold obtained using simulation.

ARL	Theory b	Simulated b
5000	7.37	7.04
10000	8.05	7.64

Proof techniques

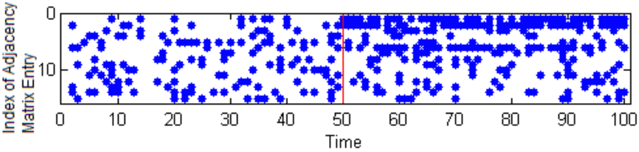
- detection statistic forms a random walk $\max_{k<t} Z_{k,t}$



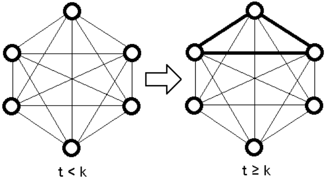
- to calculate ARL = calculate boundary hitting probability of a noise-like random walk
- the moment of detection of a stopping time, asymptotically exponentially distributed
- when $b \to \infty$, approximately $\mathbb{P}^{\infty}\{\max_{k<t} Z_{k,t} \geq b\} \approx m\lambda$

Numerical performance analysis

Detect emergence of a community



interact with $p_0 \Rightarrow$ interact with p_1



size of community is $|\mathcal{S}^*| = s$

Table : Comparison of detection delays for various cases when $N = 6$. The numbers inside the brackets are the threshold b such that ARL = 5000.

	$T_{\mathrm{ES},1}$ $\delta = p_1$	T_{Mix} $\delta = p_1$	$T_{\mathrm{H-Mix}}$	T_{Mix} $\delta = p_1 - 0.1$
$s = 3$, $p_0 = 0.2$, $p_1 = 0.9$	3.8 (9.96)	4.3 (6.71)	3.8 (9.95)	6.0 (6.71)
$s = 3$, $p_0 = 0.3$, $p_1 = 0.7$	9.5 (10.17)	12.8 (6.77)	10.8 (10.18)	23.3 (6.77)
$s = 4$, $p_0 = 0.3$, $p_1 = 0.7$	5.0 (8.48)	6.7 (6.88)	6.4 (10.17)	11.0 (6.88)

Robustness against false communities

Table : ARL and DD for each algorithm under the conditions $p_0 = 0.2, p_1 = 0.9, k = 3$, and $N = 6$ where the ARL = 5000.

	Threshold	Detection Delay
$T_{\mathrm{ES},1}$	9.96	49.74
T_{Mix}	6.71	4.30
$T_{\mathrm{H-Mix}}$	9.95	100.74

Mixture method incorrectly reacts to false community very quickly.

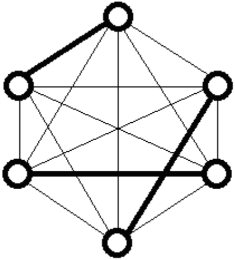
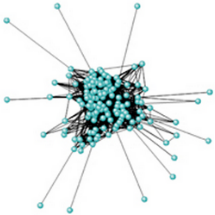


Table : Comparison of detection delays for a larger network $N = 50$. The numbers inside the brackets are the threshold b such that ARL = 5000.

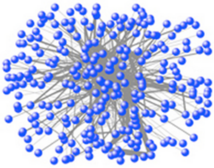
	T_{Mix}, $\delta = p_1$
$p_0 = 0.3$, $p_1 = 0.7$, $s = 10$	27.5 (-7.44)
$p_0 = 0.3$, $p_1 = 0.7$, $s = 20$	1.1 (-7.41)

Summary

- detect emergence of a community in sequential data
- present a new change-point detection approach
- three methods: exhaustive search (ES), mixture (Mix), and hierarchical mixture (H-Mix) methods, all able to detect the community quickly in different settings
  - complexity: Mix $<$ H-Mix $\ll$ ES
  - robustness: Mix $<$ H-Mix $\approx$ ES
- accurate theoretically characterize performance of Mix method
- future: apply on real data and larger networks: Enron data set



ILLICIT



LEGAL