PET: A Statistical Model for Popular Events Tracking in Social Communities

Cindy Xide Lin[1], Bo Zhao[1], Qiaozhu Mei[2], Jiawei Han[1]

[1] University of Illinois at Urbana-Champaign

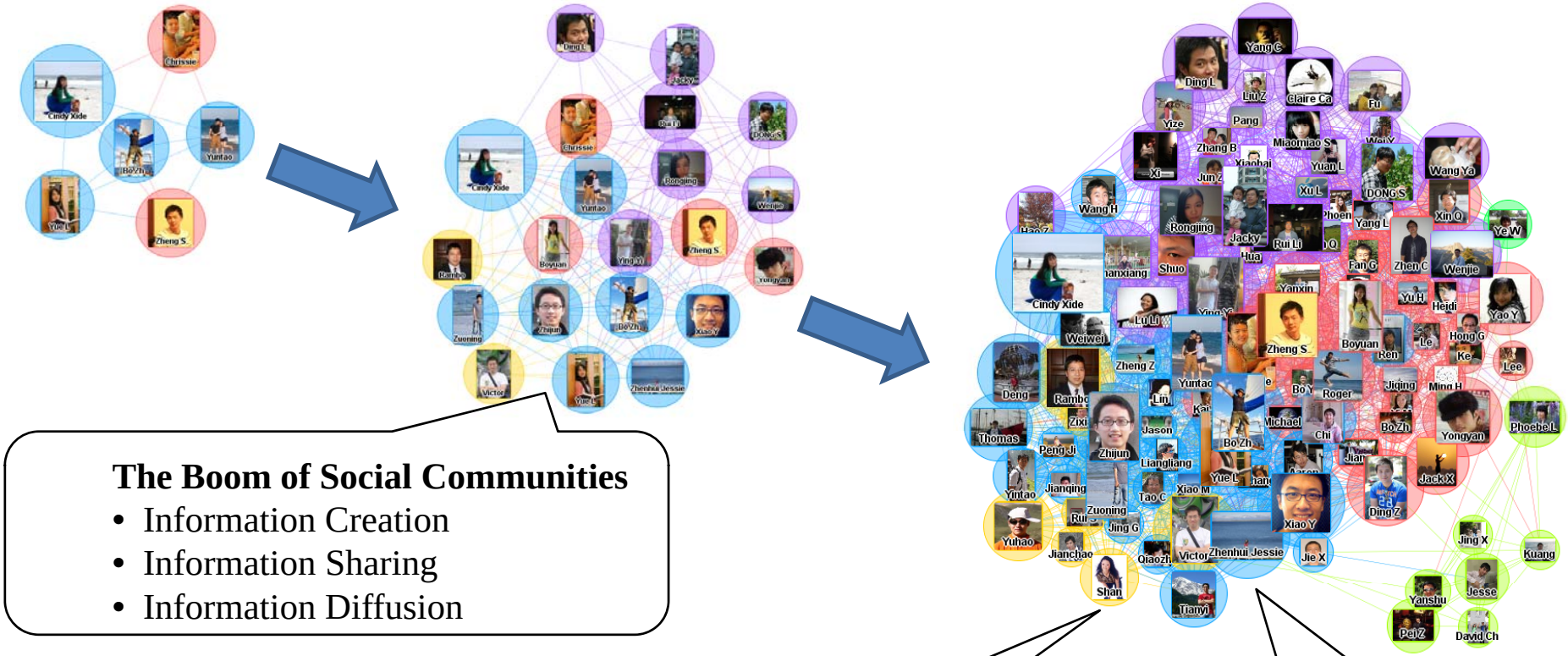
{xidelin2, bozhao3, hanj}@illinois.edu

[2] University of Michigan

qmei@umich.edu

KDD 2010

Motivation



The Boom of Social Communities

- Information Creation
- Information Sharing
- Information Diffusion

Tracking Popular Events

- Who initialized a rumor?
- Who are still interested in Avatar at 1/1/2010 ?
- What do people say about Tiger Woods before and after the scandal?

What can we utilize?

- Text
- History
- Network Structure

Motivation (Cont')

A novel probabilistic model (called **PET**) is proposed for popular events tracking in a time-variant social community that consists of both a stream of text information and a stream of network structures.

- A Gibbs Random Field models the interest of users.
- A topic model is designed to explain the generation of text data given the interest of users
- They thus interplay by regularizing each other.
- An optimization problem that considers historic, textual, and structural features.

Outline

- Motivation
- Problem Formulation
- Event Tracking Model
- Experiments
- Ongoing and Future Work
- Conclusion

Problem Formulation

Key Concepts

- Network Stream. Let $G = \{G_1, G_2, \ldots, G_T\}$ be a stream of network structures. G_k is a snapshot of a general network G at time k, where a vertex $v_{k,i}$ stands for a user and an edge $g_{k,i,j}$ corresponds to a connection between two vertices.
- Document Stream. Let $D = \{D_1, D_2, \ldots, D_T\}$ be a stream of document collections. D_k is the set of documents published at time k, where $d_{k,i} \in D_k$ is the text document(s) associated with the user $v_{k,i}$ in G_k.
- Event. A semantically coherent topic θ is a multinomial distribution of words $\{p(w|\theta)\}_w$. We define an event as a stream of topics $\Theta = \{\theta_0, \theta_1, \ldots, \theta_T\}$, where θ_0 is the primitive topic describing the event, and θ_k corresponds to the version of θ_0 at time k.
- Interest. For a particular event, at each time point k, we assume each user has a certain level of interest in that event. We model such level of interest as a real value $h_{k,i} \in [0, 1]$, and denote the set of interest values for all vertices in G_k as H_k.

Problem Formulation (Cont')

- **Input**
  - Primitive Topic θ_0 that describes a popular event
  - An Observed Stream of Networks $G = \{G_1, G_2, \dots, G_T\}$
  - An Observed Stream of Documents $D = \{D_1, D_2, \dots, D_T\}$

- **Output**
  - A Latent Stream of Interest $H = \{H_1, H_2, \dots, H_T\}$
  - A Latent Stream of Topics $\Theta = \{\theta_1, \theta_2, \dots, \theta_T\}$

Outline

- Motivation
- Problem Formulation
- Event Tracking Model
- Experiments
- Ongoing and Future Work
- Conclusion

Event Tracking Model

- **Intuitions**
  - Observation 1: the interaction between interests H_k and connections G_k.
  - Observation 2: the interaction between interests H_k and history H_{k-1}.
  - Observation 3: the interaction between contents D_k and interests H_k.
- **Independence Assumptions**
  - Assumption 1: given the current network structure G_k and the previous interest status H_{k-1}, the current interest status H_k is independent of the document collection D_k.
  - Assumption 2: given the current interest status H_k and the document collection D_k, the current topic model θ_k is independent of the network structure G_k.

Event Tracking Model

Formally, the task of PET is cast as the inference of the posterior of H_k and θ_k.

Markovian Simplification

$$P(H_k, \Theta_k | G_k, D_k, H_{k-1}) =$$
$$P(H_k | G_k, H_{k-1}) \cdot P(\Theta_k | H_k, D_k)$$

Interest Model, based on
- Observation 1 & 2
- Assumption 1

Topic Model, based on
- Observation 3
- Assumption 2

The Interest Model

We propose a multivariate Gibbs Random Field to model the dependency among users and the influence of past status. Formally, the interest status H_k is a family of random variables defined on graph G_k, and we give a configuration of H_k that follows a Gibbs distribution:

Partition Function $Z = \sum_{f \in \mathbb{F}} P(f.)$

Temperature λ_T

$$P(H_k|G_k, H_{k-1}) = Z^{-1} \times e^{-\frac{1}{\lambda_T} U(H_k)}$$

Energy Function

$$U(H_k) = \sum_{i=1}^{N} V_i(h_k(i)) + \sum_{i=1}^{N} V_i'(h_k(i), h_k(-i))$$

- Potential Function 1: models the transition energy.

$$V_i(h_k(i)) = (h_k(i) - h_{k-1}(i))^2, \forall i \in [1..N]$$

Observation 2

- Potential Function 2: gives penalty for the difference.

$$V_i'(h_k(i), h_k(-i)) = \lambda_{k,i}(h_k(i) - h_k'(i)))^2, \forall i \in [1..N]$$

Observation 1

The Topic Model

We consider each document $d_{k,i} \in D_k$ is generated from the mixture of two models $\Theta_k = \{\theta_B, \theta_k\}$, where θ_B is the background model and θ_k is the latent topic model that we want to estimate. Formally, the posterior of topic Θ_k is given as

The likelihood of the document collection D_k

$$P(D_{k,i}|H_k, \Theta_k) \propto \prod_{i=1}^{N} \prod_{w \in W} p(w|d_{k,i})^{c(d_{k,i},w)}$$

The probability of generating word w in $d_{k,i}$ is

$$p(w|d_{k,i}) = h_k(i)p(w|\theta_k^E) + (1 - h_k(i))p(w|\theta_k^B)$$

Observation 3

$$P(\Theta_k|H_k, D_k) \propto P(D_k|H_k, \Theta_k)P(\Theta_k|H_k)$$

Dirichlet Prior:

$$P(\Theta_k|H_k) = P(\theta_k^E) \propto \prod_{w \in W} p(w|\theta_k^E)^{\mu_E p(w|\theta_0^E)}$$

Parameter Estimation

Given our model defined as above, we can fit the model to the data and estimate the parameters by the Expectation Maximization (EM) algorithm.

In the E-Step, we compute the expectation of the hidden variables as

$$p^{(n)}(z_{d_{k,i},w} = \theta_k^E) = \frac{h_k^{(n-1)}(i)p^{(n-1)}(w|\theta_k^E)}{h_k^{(n-1)}(i)p^{(n-1)}(w|\theta_k^E) + (1 - h_k^{(n-1)}(i))p^{(n-1)}(w|\theta_k^B)}$$

In the M-Step, the object function that we want to maximize is

$$-\log Z - \frac{1}{\lambda_T}\sum_{i=1}^{N}\left((h_k(i) - h_{k-1}(i))^2 + \lambda_{k,i}(h_k(i) - h'_k(i))^2\right)$$

$$+\sum_{i=1}^{N}\sum_{w\in W} c(d_{k,i}, w)p^{(n)}(z_{d_{k,i},w} = \theta_k^E)\log(h_k(i)p(w|\theta_k^E))$$

$$+\sum_{i=1}^{N}\sum_{w\in W} c(d_{k,i}, w)p^{(n)}(z_{d_{k,i},w} = \theta_k^B)\log((1 - h_k(i))p(w|\theta_k^B))$$

$$+\sum_{w\in W}\mu_E p(w|\theta_0^E)\log(p(w|\theta_k^E))$$

Interest Model

Component Model θ_k

Component Model θ_B

Dirichlet Prior

Parameter Estimation (Cont')

By integrating Lagrange multipliers, the inference of $h_{k,i}$ boils down to solve:

$$\alpha h_k(i) - \beta - \frac{\gamma}{h_k(i)} - \frac{\delta}{h_k(i)-1} = 0$$

where

$$\alpha = \frac{2}{\lambda_T}(1+\lambda_{k,i}),$$

$$\beta = \frac{2}{\lambda_T}(h_{k-1}(i) + \lambda_{k,i}h'_k(i)),$$

$$\gamma = \sum_{w \in W} c(d_{k,i}, w) p^{(n)}(z_{d_{k,i},w} = \theta_k^E),$$

$$\delta = \sum_{w \in W} c(d_{k,i}, w) p^{(n)}(z_{d_{k,i},w} = \theta_k^B).$$

In the case of $\sum_{w \in W} c(d_{k,i}, w) = 0$, $h_{k,i}$ only depends on its past and neighborhood:

$$h_k(i) = \frac{\beta}{\alpha} = \frac{h_{k-1}(i) + \lambda_{k,i}h'_k(i)}{1+\lambda_{k,i}}$$

In the case of $\sum_{w \in W} c(d_{k,i}, w) > 0$, it is equivalent to solve the cubic function:

$$\alpha h_k(i)^3 - (\alpha+\beta)h_k(i)^2 + (\beta-\gamma-\delta)h_k(i) + \gamma = 0$$

Outline

- Motivation
- Problem Formulation
- Event Tracking Model
- Experiments
- Ongoing and Future Work
- Conclusion

Experiments

Dataset

- Twitter: a free social networking and micro-blogging service.
  - Users can send and read messages known as tweets.
  - Users have follower-followee relationships.
  - *50,000* users with *1, 438, 826* tweets displayed from *Oct. 2009* to *Dec. 2009*
  - The document $d_{k,i}$ is the concatenation of tweets published by user i at day k.
  - The network structure G_k is generated according the replying frequency among users during the past *30* days.

- DBLP: a database that contains the basic bibliographic information of computer science publications.
  - *2,949* authors who published at least 10 papers in area of data mining and database
  - *500,417* papers during the period from *1990* to *2008*.
  - The document $d_{k,i}$ is the concatenation of paper titles published by user i at year k.
  - The network structure G_k is the co-author network at year k.

Experiments

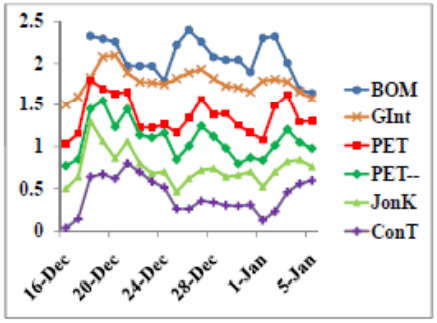
Baseline and Ground Truth

- JonK: a state automation model based on HMM.
  - It is a variation of Jon Kleinberg's model in the literature '*bursty and hierarchical structure in streams, KDD2002*'.
  - It is a special case of the PET model when we make certain constraints and simplification to remove the network effects.
- ConT: a contagion model
  - It is introduced by S. Morris in the literature '*Contagion. In Review of Economic Studies, 2000*'.
  - It is a special case of the PET model when the topic effects are omitted.
- PET-: a special version of PET
  - To show the contribution made by network structures for popular event tracking
  - It is implemented by removing network structures from PET, i.e., setting $g_{k,i,j} = 0$.
- BOM: the ground truth for movie-related events
  - The daily box office at *http://boxofficemojo.com/movies*.
- GInt: the ground truth for news-related events
  - The the interest index by analyzing the search volume of Google at *http://www.google.com/insights/search.*

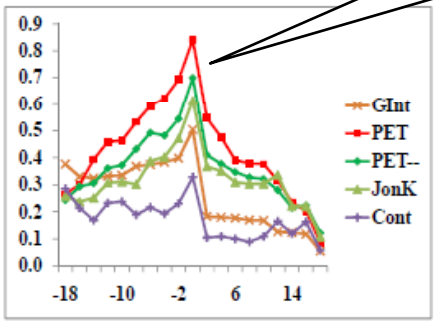
Experiments

Analysis on Popularity Trend

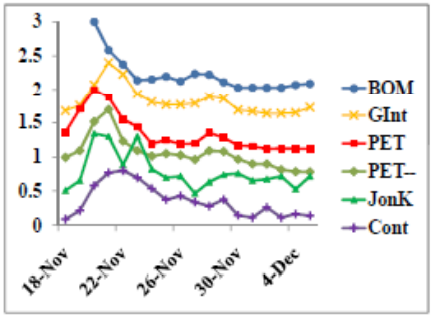
PET (the red one) has the highest cross-correlation score with the ground truth.



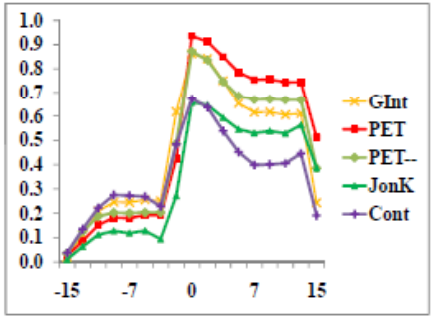
(a) the popularity evolution on 'avatar'



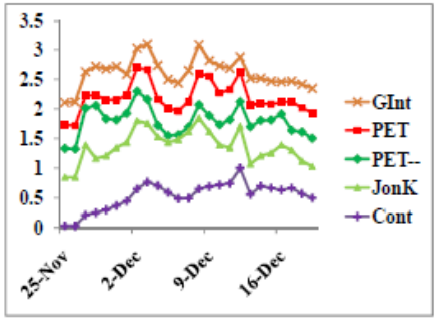
(b) the correlation analysis on 'avatar'



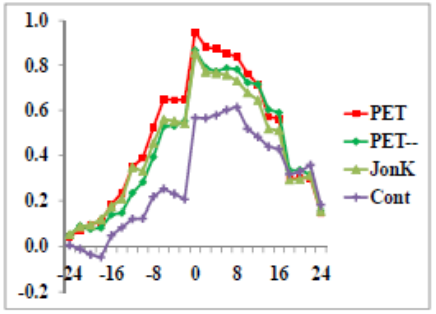
(c) the popularity evolution on 'twilight'



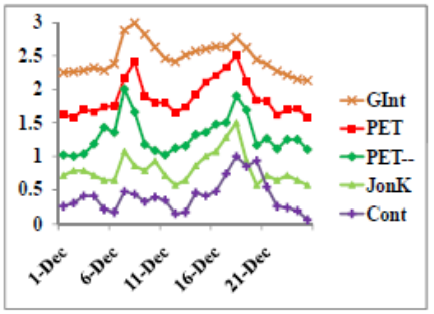
(d) the correlation analysis on 'twilight'



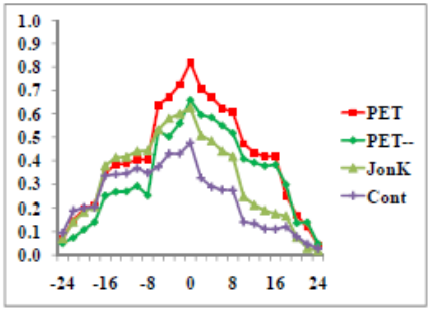
(e) the popularity evolution on 'Tiger Woods Affair'



(f) the correlation analysis on 'Tiger Woods Affair'



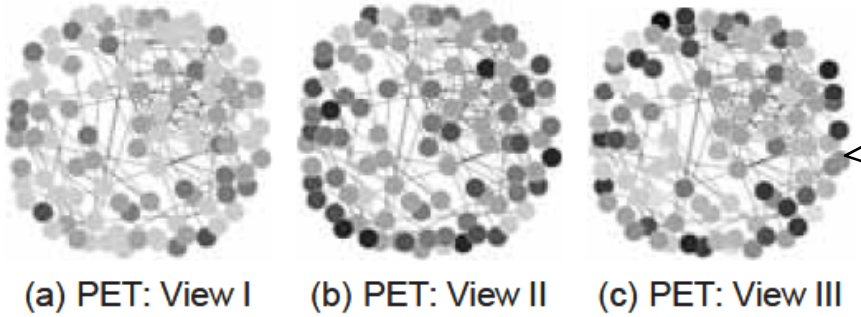
(g) the popularity evolution on 'Copenhagen'



(h) the correlation analysis on 'Copenhagen'

Experiments

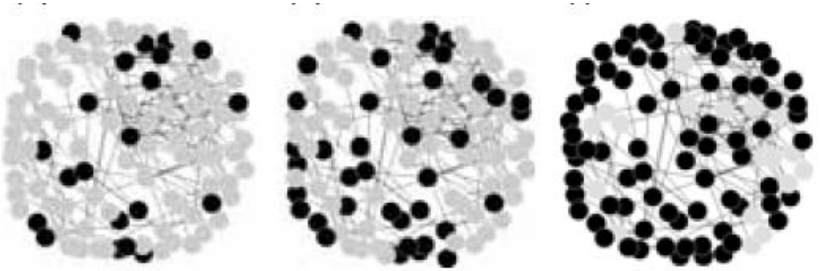
- **Analysis on Network Diffusion**



(a) PET: View I (b) PET: View II (c) PET: View III

(d) PET–: View I (e) PET–: View II (f) PET–: View III

Vertices in networks of PET are smoothed via edges, which accords better with the situation in real world: people's interests are inevitably influenced by their friends.



(g) Cont: View I (h) Cont: View II (i) Cont: View III

Experiments

- **Analysis on Content Evolution**

Movie Contents

Movie Title, Director, Actors

Historical Significance

Dec 14		Dec 18		Dec 26	
trailer	0.21	avatar	.30	avatar	0.13
avatar	0.10	imax	0.06	imax	0.04
cameron	0.04	trailer	0.05	trailer	0.04
james	0.02	cameron	0.04	technology	0.03
sam	0.01	james	0.04	sam	0.02
director	0.01	alien	0.01	film	0.02
titanic	0.01	titanic	0.01	james	0.02

Car Accident, Competition

Limited Information

Nov 25		Nov 27		Dec 10	
rhapsody	0.07	tiger	0.11	tiger	0.10
muppets	0.06	woods	0.09	woods	0.06
bohemian	0.06	injure	0.04	brown	0.02
lambert	0.01	car	0.03	mistress	0.01
tiger	0.01	accident	0.03	golf	0.01
woods	0.01	championship	0.01	shame	0.01
playlist	0.01	hospital	0.01	divorce	0.01

Sex Scandal, Marriage

Outline

- Motivation
- Problem Formulation
- Event Tracking Model
- Experiments
- Ongoing and Future Work
- Conclusion

Ongoing and future work

- **Can we track a set of popular events simultaneously ?**
  - The background model could be a mixture of other popular events.
  - The efficiency issue
- **What are the challenges ?**
  - The number of popular events is almost infinite.
  - A user is not interested in all events.
- **Personalized popular events tracking**
  - The number of personalized popular events is reasonably small.
  - Even personalized information could be huge amount – We need highlights and summarization.
  - Help create more social connections.

Ongoing and future work

Wall Photos
Feed team says thanks to our 500 million users
By: Jing Chen

Yongyan Liu and **Rui Li** were tagged in a photo.

Profile Pictures

Cathy Wu yay 500 million active facebook users yay
on Wednesday · Comment · Like · Share

Yefei Wang do you have to both start and end with a yay?
Wednesday at 4:58pm · Like · 1 person

Cathy Wu yes, yes I do
Wednesday at 5:10pm · Like

All about the same event:

reached 500 million users.

Conclusion

In this work, we propose the novel problem of popular events tracking in a social community. Given a stream of network structures, an associated stream of text documents, and the primitive form of events, we could track the popularity of the events on the network and content revolution of the events over time.

The proposed model, PET, not only provides a unified probabilistic framework to model different factors in modeling the evolution of interests and contents, but also covers classical models as special cases.

Thank you

May, 2010, Urbana-Champaign