Review Spam Detection

Bimodal Distribution and Co-Bursting

H. Li, G. Fei, S. Wang, B. Liu, W. Shao, A. Mukherjee, and J. Shao. Bimodal distribution and co-bursting in review spam detection
In International Conference on World Wide Web, 2017

Motivation

Who cares about online reviews?

- Commercial interest
  - Direct profit
  - Trust
  - Design decisions
- Academic interest
  - **Untruth**ful Reviews
- Is it a big deal?
  - Estimated that a third of all online reviews are fake

Presentation Goal

Introducing a new, cutting edge, method to identify fake reviews

Structure

- Introduction
    - Devil's advocate
    - Knows methods and challenges
- The data set
- Reviews as a time series
    - The frequency of reviewing
- Using HMM (The model)
- Using LHMM (The Full model)
    - Why should the model work
- What is HMM?
- Co-bursting of reviewers
- Results and conclusions

Introduction

If I were a spammer...

- Use similar words, sentence structure, etc.
- Mind your reviews frequency
  - Be active, but not too active
- Don't be an outlier, create a trend
  - Work with friends
- Be your fake account
  - Have different accounts for different “characters”

Why is this hard?

- Even for humans, it is very hard to know when a review is fake
  - Very limited training sets
  - Imbalanced class distribution
- Many different domains
  - And languages
- Spamming in groups
- But
  - Spamming still needs to be cheap

Find the fake

Great Hotel This building has been fantastically converted into studios/suites. We only had a studio which was brilliant can’t imagine how the suite could have bettered what we had. The kitchen had everything cooker microwave dishwasher and fridge freezer…..

During my latest business trip, both me and my wife recently stayed at the Omni Chicago Hotel in Chicago, Illinois, at one of their Deluxe suites. Unfortunately, and I think I speak for both of us, we were not fully satisfied with the hotel. The hotel advertises luxury-level accommodations, and while the rooms resemble what one can see in the pictures, the service is certainly sub-par. When one plans a stay at such an establishment, they expect a service that goes beyond having fresh towels in the bathroom when they check in……..

Known methods for spam review detection

Using the review

- Bag of words, n-grams, term frequency
- Part of speech tagging
- Lexical features (Average word length)
- Syntactic features (Number of function words)
- Semantic features
- Metadata - length, date, time, rating, etc.

Using the reviewer

- Profile characteristics
- Behavioral patterns - RPD, positive rate, length, rating distribution
- Maximum content similarity
- Reviewer - product networks

The main idea in this new work:
Look at the review
POSTING TIME

The data set

Meituan- Dianping

(One of) The biggest companies in the world you never heard about

About Dianping

- The biggest (restaurant) review site in the world
- In 2015 merged with Meituan to become the largest online and on-demand delivery platform
- Over 180 million monthly active users, 600 million registered users, almost 4.5 million business partners

Used in this research

- 1.5M **labeled** reviews, from 68K reviewers
  - Only reviewers with more than 10 reviews
  - Starting from 2.7M reviews from 633K reviewers
- All published reviews for any of the reviewers
- Labels from Dianping commercial spam filter

The frequency of reviewing

What is the distribution of review time intervals?

- It’s a Poisson Process right?
  - Events occur continuously and independently at a constant average rate
- Wrong!

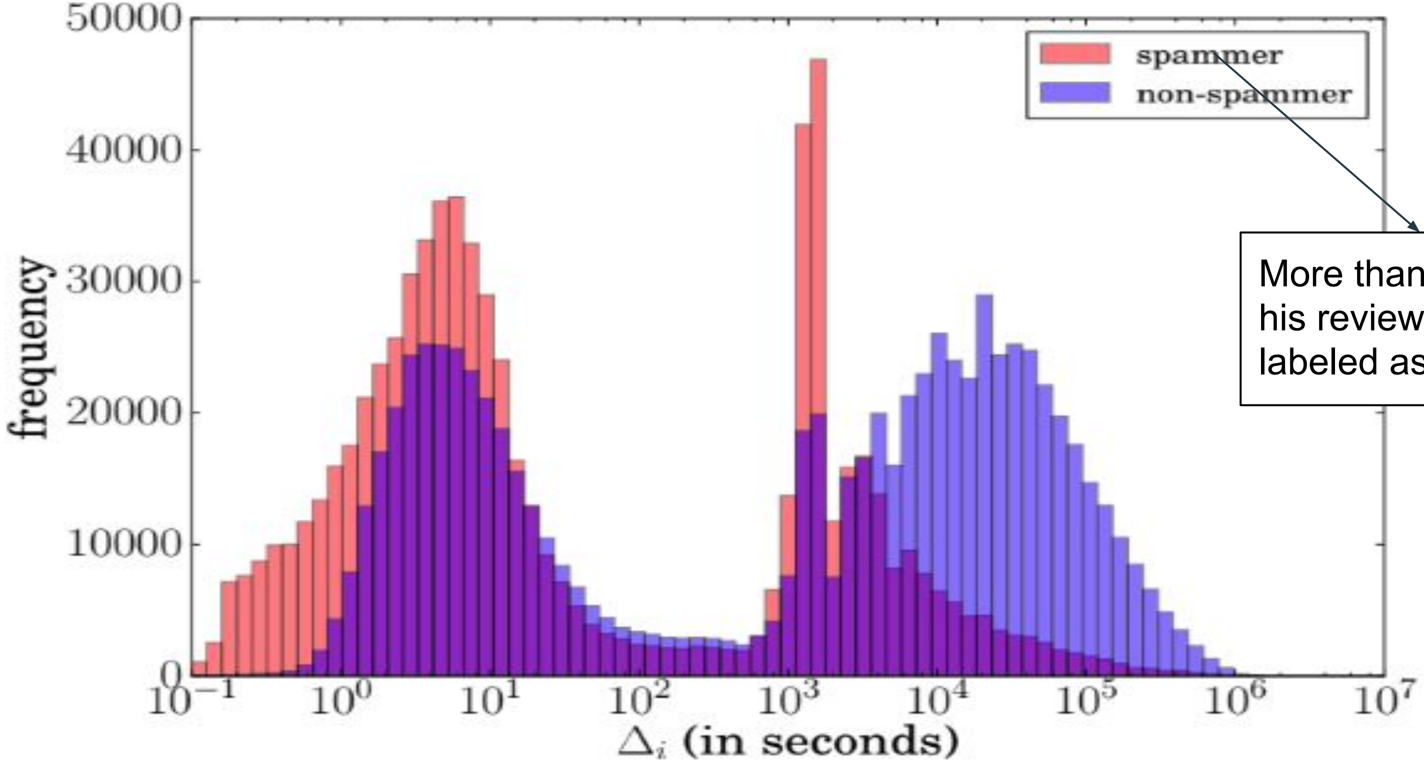
$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geqslant 0 \\ 0 & x < 0 \end{cases}$$

$$E(X) = \frac{1}{\lambda}$$

$$\boxed{\Delta_i \triangleq t_i - t_{i-1}}$$

$$\Delta_i \sim Exp(\lambda)$$

Bimodal distribution of time intervals between adjacent reviews



More than 10% of his reviews are labeled as spam

What are we seeing?

- For both spammers and non spammers, writing a review is “a process with memory”
- Both follow a bimodal distribution with very distinct, separated peaks

Non Spammers

- Have the tendency to write a few reviews after a period of inaction to summarize their recent experiences after eating in some restaurants.
- Much longer tail
- The mean is 2-3 times longer

Spammers

- Participate in spam attacks/campaigns and write many reviews during a campaign but do not write much before or after that
- Many reviews with short intervals

So what?

- So, first, users will be classified into one of two states: Active/Fast and Inactive/Slow
- Then, the distribution and transition function between states will be used to identify spammers
- Also, spammers work in groups, this will be used to **help** identify them

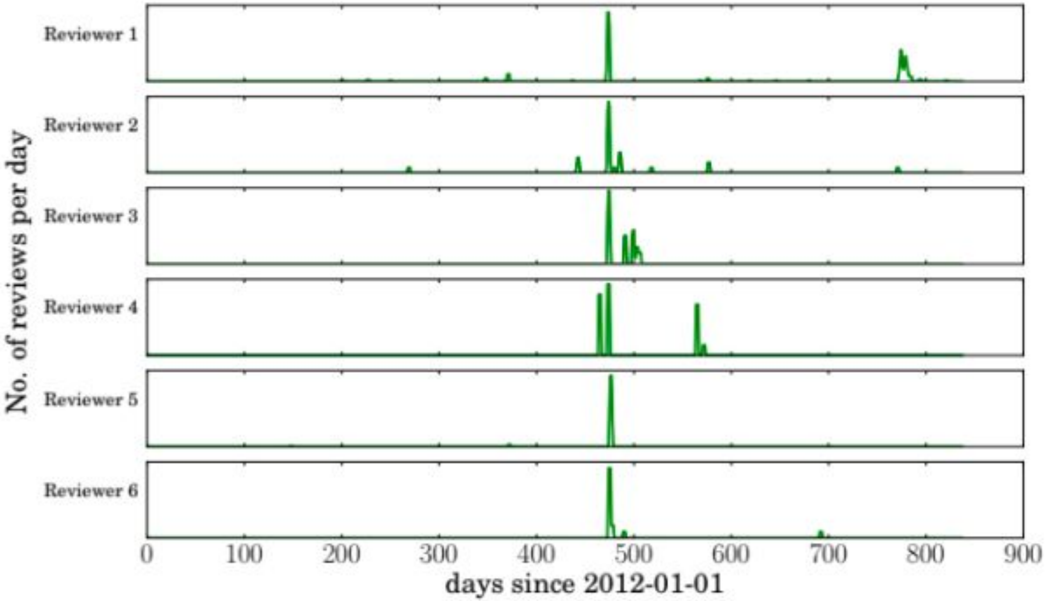


Figure 1: Examples of co-bursting behaviors

Using HMM

How do we model this?

Using A Hidden Markov Model (HMM)

- A Markov Model (First order) has a memory of one state
  - Makes sense? Short-term memory of human behaviors
  - Works well for reviews - Found strong correlations between consecutive time intervals
- In a HMM the states are hidden, we only observed signals (Δ_i) emitted from the hidden states
- Δ_i may follow **different** exponential distributions depending on state
- The point is to model the transitions between Δ_i for every reviewer and to solve the decoding problem which aims to estimate the most likely state sequence in the model, given the observations

Mathematical description

$$\mathit{States} = \{Q_0,\ Q_1\}$$

$$\mathit{Emission\ parameter} = \Delta_i$$

$$\mathbb{A} = \{a_{kj}\} = \begin{bmatrix} a_{0,0} & a_{0,1} \\ a_{1,0} & a_{1,1} \end{bmatrix}$$

$$a_{kj} = P(Q_i = j | Q_{i-1} = k),\ k, j \in \{0, 1\}$$

Probability matrix between states

$$\Delta_i \sim \begin{cases} Exp(\lambda_0), & Q_i = 0 \\ Exp(\lambda_1), & Q_i = 1 \end{cases}$$

$$\mathbb{B} = \{b_j(\Delta)\} \text{ and } b_j(\Delta) = f(\Delta; \lambda_j) = \lambda_j e^{-\lambda_j \Delta}$$

Probability distribution of an observation

$$P(Q_{1:T}, \Delta_{1:T})$$

$$= P(Q_1, Q_2, \Delta_2, \ldots, Q_T, \Delta_T)$$

$$= P(Q_1) \prod_{i=2}^{T} P(\Delta_i | Q_i) \prod_{i=2}^{T} P(Q_i | Q_{i-1})$$

Joint probability of the observations and hidden states

To Identifying the state sequence for each reviewer, this needs to be maximized

$$Q^*_{1:T} = \underset{Q_{1:T}}{\operatorname{argmax}} P(Q_{1:T} | \Delta_{1:T})$$

Using LHMM

Aren’t we missing something?

Oh yes, the spam part… We need Labeled Hidden Markov Model

- LHMM is the “novelty” of the paper
- The idea is to introduce a new binary variable Y to represent the labels
  - Y = + stands for spammers and Y = - for non-spammers
- Now the states transition probability matrix and the probability distribution of the observations will be dependent on the reviewer class
- In order to predict the value of Y given the observations, the Bayesian theorem is needed
- The most probable value that the class variable takes is the one that better explains or generates the observations

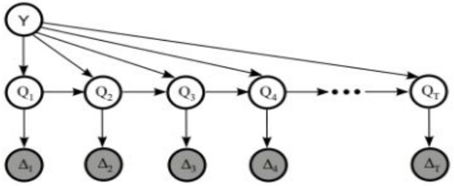


Figure 4: Representation of Labeled Hidden Markov Model

The Heart of LHMM

- All the presented math changes accordingly, for example:

$$\Delta_i \sim \begin{cases} Exp(\lambda_0^Y), & Q_i = 0 \\ Exp(\lambda_1^Y), & Q_i = 1 \end{cases}$$

- Now we need to find the most probable Y for each reviewer

$$y^* = \underset{y}{\operatorname{argmax}}\, P(Y = y|\Delta_{1:T})$$

- Using Bayesian theorem:
  - The denominator is independent of Y - can be dropped
  - The second factor in the numerator is easily calculated

$$= \underset{y}{\operatorname{argmax}} \frac{P(\Delta_{1:T}|Y = y) \cdot P(Y = y)}{P(\Delta_{1:T})}$$

- The first factor will be calculated using our previously deduced joint probability of observations and hidden states
- The calculation itself will be implemented using a dynamic programing algorithm named “forward-backward method”, in almost linear time.

$$\begin{aligned} &P(\Delta_{1:T}|Y) \\ &= \sum_{Q_{1:T}} P(Q_{1:T}, \Delta_{1:T}|Y) \\ &= \sum_{Q_{1:T}} P(Q_1|Y) \prod_{i=2}^{T} P(\Delta_i|Q_i, Y) \prod_{i=2}^{T} P(Q_i|Q_{i-1}, Y) \end{aligned}$$

Now that we have

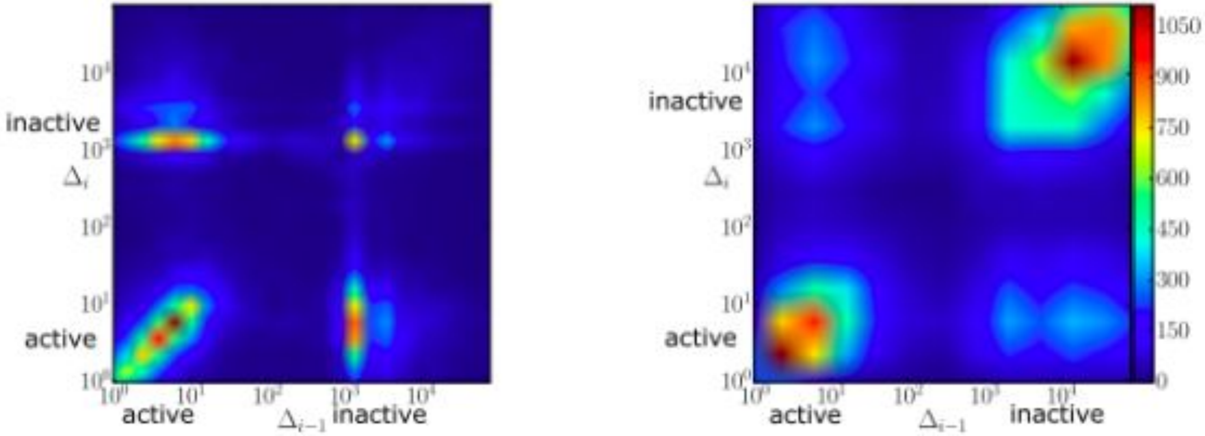
$$y^* = \underset{y}{\operatorname{argmax}}\, P(Y = y|\Delta_{1:T})$$

We are done!

Why would LHMM work

But wait, why is this supposed to work again?

Are the time interval series really that different for spammers and non-spammers?



(a) spammers

(b) non-spammers

Figure 3: Heatmap of consecutive time interval pairs (in seconds). Each point corresponds to (Δ_{i-1}, Δ_i) for some reviewer.

The difference in the transition patterns

- The heat map represents the second derivative
- It has 4 regions corresponding to the 4 main transitions: Fast-Fast, Fast-Slow, Slow-Fast, Slow-Slow
- Showing the difference in transition between spammers and non-spammers will solidify our understanding of the validity of using LHMM

Active To Active

Spammers

- Strong correlation
- Active campaign posts are like a job - constant fast rate with high self discipline

Non Spammers

- Weak correlation
- Even in active state regular users don’t have a constant rate

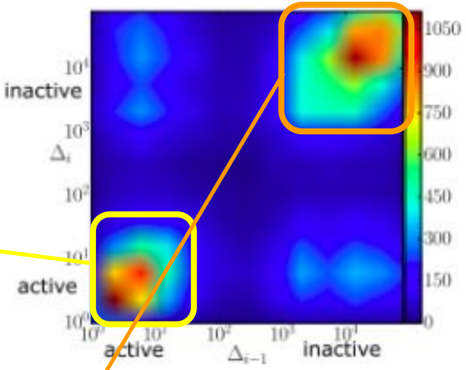
Inactive To Inactive

Spammers

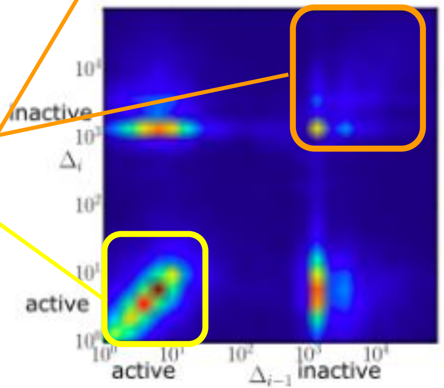
- Rest in a very similar way
- Almost no activity except one “long term” pattern

Non Spammers

- Very Weak correlation
- Rest very differently, many have sporadic posts



(b) non-spammers



(a) spammers

Inactive To Active

Spammers

- When activated, they start working in a variety of different rhythms depending on their campaign needs

Non Spammers

- When activated, their activity rates are similar (The typical “Human” activation rate)

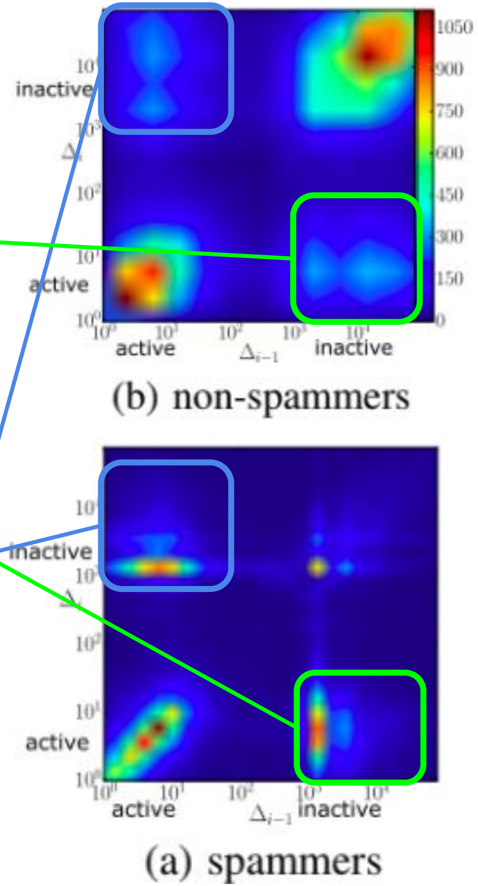
Active To Inactive

Spammers

- When deactivated, they are very similar - the campaign is over, nothing to do

Non Spammers

- “Hibernate” differently due to their own habits
- The time it takes to write a review after writing the last active review is significantly different



(b) non-spammers

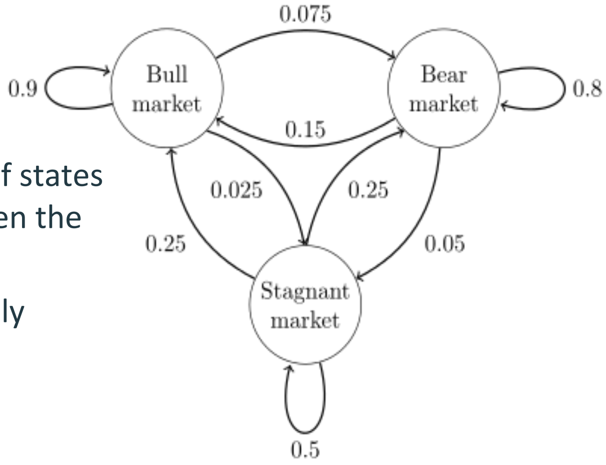
(a) spammers

Based on the discovery of the major differences between emission probability and transition probability of the two classes - LHMM works

Hidden Markov Model

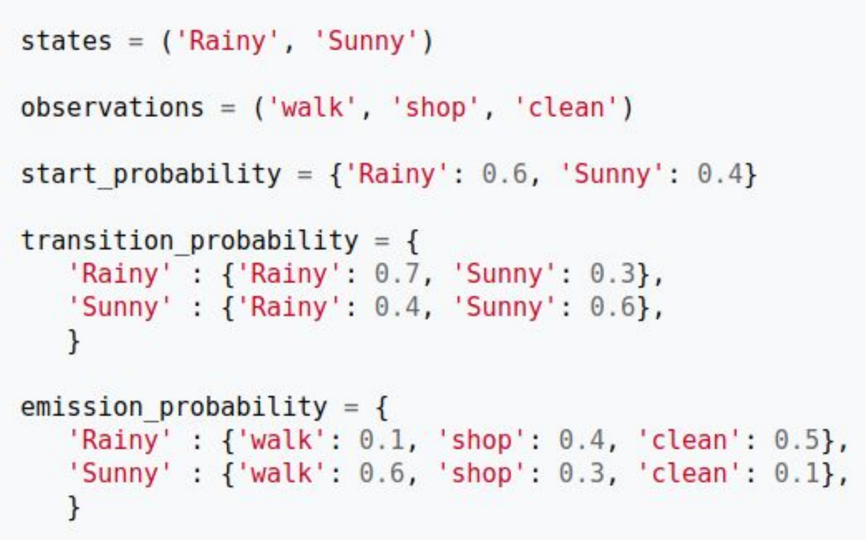
So what is a HMM anyway?

- A Markov Model is a stochastic model used to model randomly-changing systems
- It is assumed that future states depend only on the current state
  - *i.e*. has the Markov property
  - A “better memory cousin” of the Poisson process
- A simple Markov Model, *i.e*. a Markov chain, is by definition a set of states and a transition matrix that defines the probability to move between the states
- A Hidden Markov model is a Markov chain for which the state is only partially observable
  - The simplest dynamic Bayesian network
- Defined by:
  - A set of states
  - A series observations
  - Probability of transition between states
  - Probability distribution of an observation (dependent on the states)
  - Starting conditions



An Example

Consider Alice and Bob, who live far apart and talk daily over the telephone. Bob is only interested in three activities: walking in the park, shopping, and cleaning his apartment. The choice of what to do is determined exclusively by the weather on a given day. Based on what Bob tells her he did each day, Alice tries to guess what the weather must have been like.



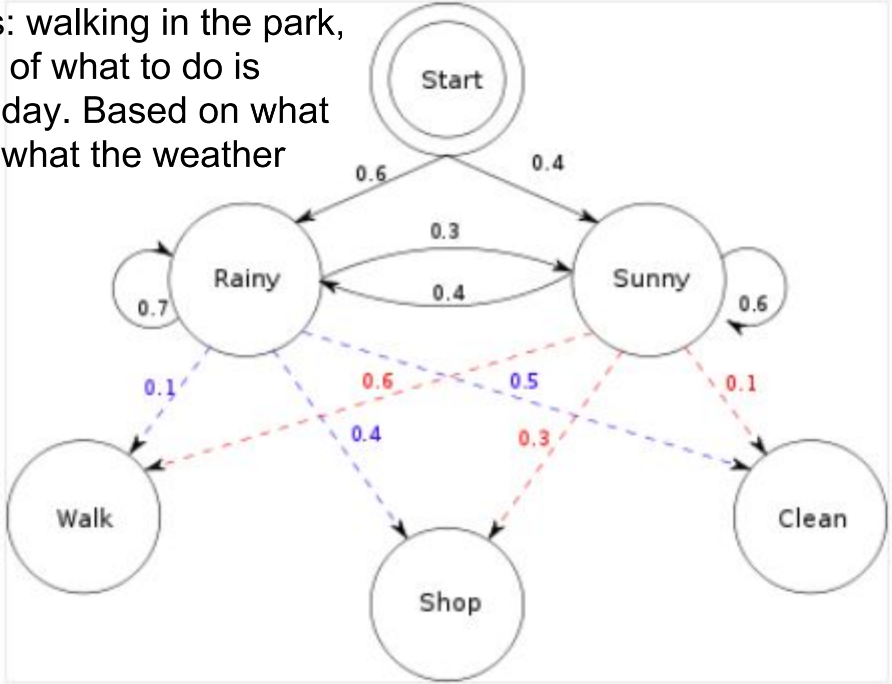
```
states = ('Rainy', 'Sunny')

observations = ('walk', 'shop', 'clean')

start_probability = {'Rainy': 0.6, 'Sunny': 0.4}

transition_probability = {
   'Rainy' : {'Rainy': 0.7, 'Sunny': 0.3},
   'Sunny' : {'Rainy': 0.4, 'Sunny': 0.6},
   }

emission_probability = {
   'Rainy' : {'walk': 0.1, 'shop': 0.4, 'clean': 0.5},
   'Sunny' : {'walk': 0.6, 'shop': 0.3, 'clean': 0.1},
   }
```



Co-bursting

Spamming with friends

- Spammers exhibit a group behavior, when one spammer bursts into action that probably means he is not the only one.
- Indreducing “Co-Bursting”
  - A group of reviewers who have bursty reviews, some of which are posted to the same set of restaurants in a short period of time
- How to measure Co-Bursting?
- With respect to a specific review **at time t** to a restaurant S from **a certain reviewer**, we consider 6 intuitive co-bursting metrics to quantify **co-spamming activities from other reviewers** who happen to write reviews to the same business within **a time window** [t -w , t+w)

1. No. of co-reviews: Counts the number of reviews of other reviewers’ to the same restaurant
2. No. of spam co-reviews: This metric is similar to the first one except that only spam reviews are counted
3. No. of co-reviews when restaurant is active: Similar to the first one except that it is conditioned on whether the restaurant of interest has bursty reviews
4. No. of spam co-reviews when restaurant is active: Similarly to previous, but only spam reviews are included
5. No. of co-reviews when reviewer is active: Similar to the first metric, this one only counts co-reviews when their **reviewers** are in the active state.
6. No. of spam co-reviews when reviewer is active: This metric considers only spam co-reviews from active reviewers.

Are those Co-Bursting metricses any good?

- Assuming each of the 6 metrics of a review is generated from a Multivariate Gaussian distribution of two set of parameters corresponding to the two different modes.

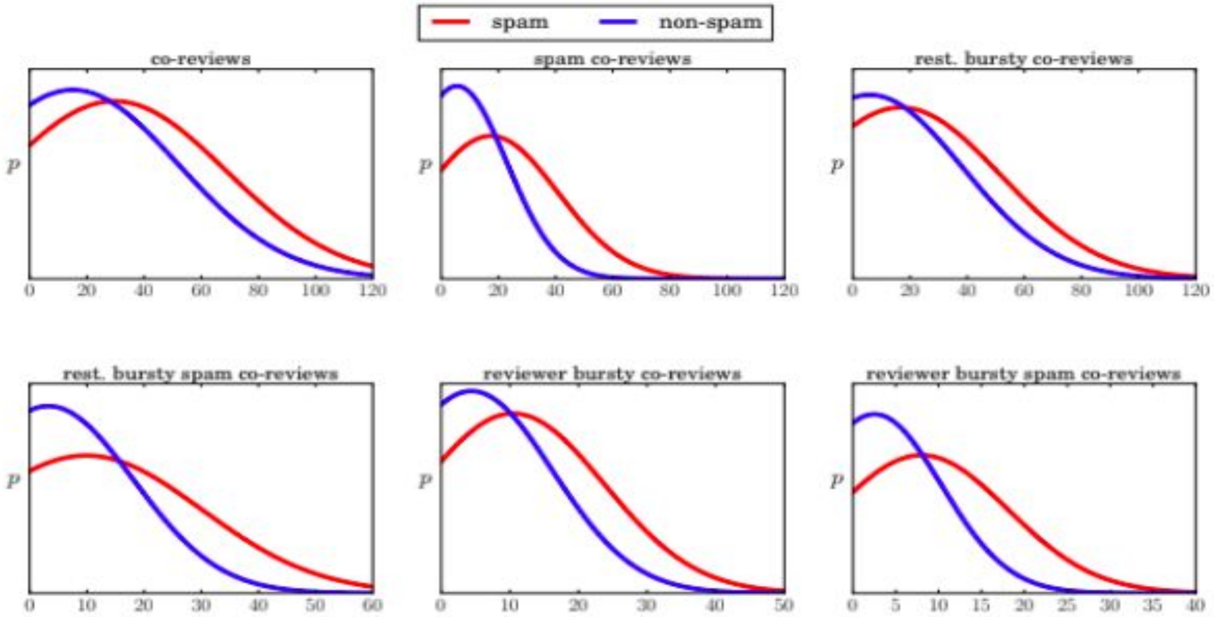


Figure 5: PDF of Gaussian distribution of co-bursting features

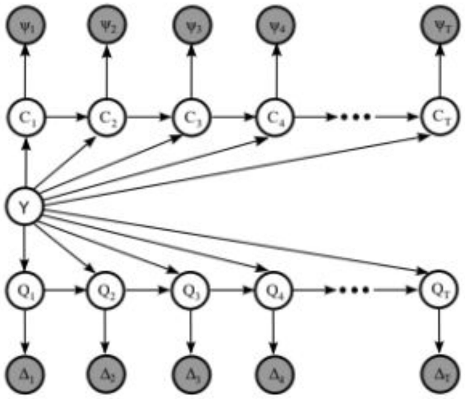
What to do with Co-Bursting?

Coupled (Labeled) Hidden Markov Model

- Extend the LHMM model to incorporate co-bursting relations to better model reviewers' collective behaviors.

- Observed co-bursting signals at t are denoted as Ψ_t which is generated from the underlying Gaussian distribution at mode C_t where $C_t \in \{0, 1\}$. $C_t = 1$ means the co-bursting mode

- Under such a framework, the inference problem becomes finding the best reviewer label Y that maximizes the joint probability with observed intervals and co-bursting signals

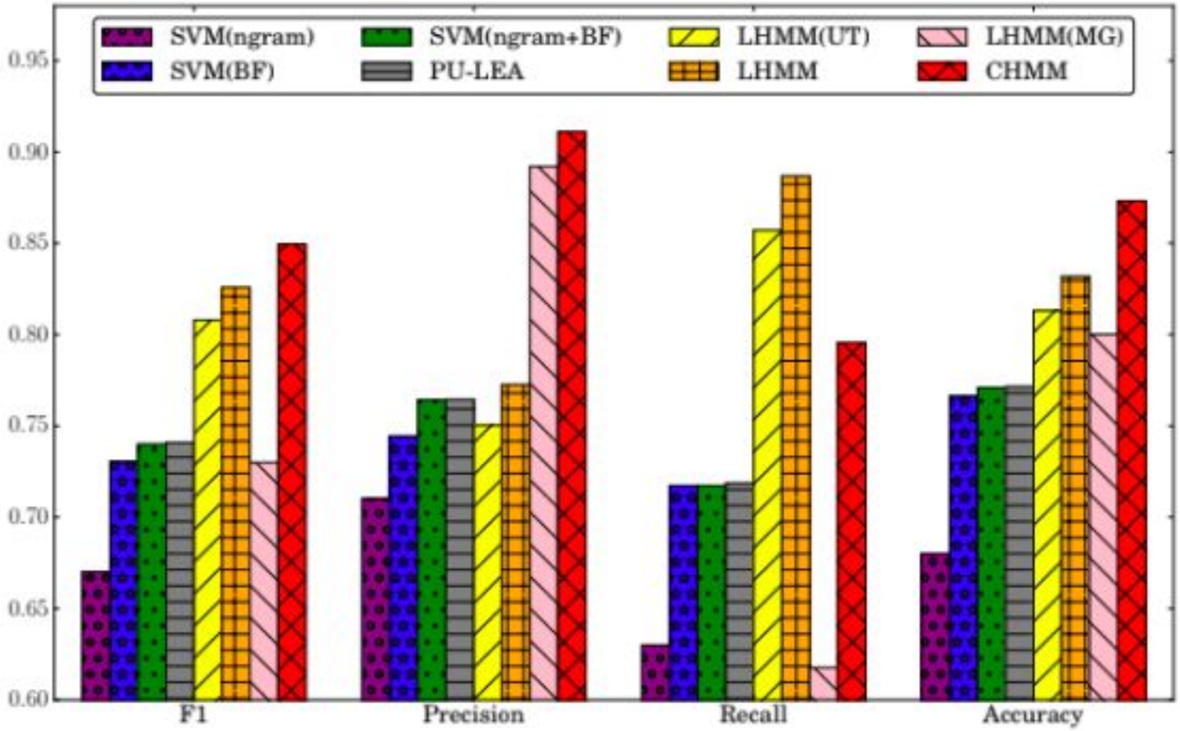


$$
\begin{aligned}
y^* &= \operatorname*{argmax}_y P(Y = y|\Delta_{1:T}, \Psi_{1:T}) = \operatorname*{argmax}_y P(\Delta_{1:T}, \Psi_{1:T}, Y = y) \\
&= \operatorname*{argmax}_y P(\Delta_{1:T}|y) \cdot P(\Psi_{1:T}|y) \cdot P(y) \\
&= \operatorname*{argmax}_y \sum_{Q_{1:T}} P(Q_{1:T}, \Delta_{1:T}|y) \sum_{C_{1:T}} P(C_{1:T}, \Psi_{1:T}|y) \cdot P(y) \\
&= \sum_{Q_{1:T}} P(Q_1|Y) \prod_{i=2}^{T} P(\Delta_i|Q_i, Y) P(Q_i|Q_{i-1}, Y) \\
&\quad \cdot \sum_{C_{1:T}} P(C_1|Y) \prod_{i=2}^{T} P(\Psi_i|C_i, Y) P(C_i|C_{i-1}, Y)
\end{aligned}
$$

DETECTING SPAMMER GROUPS

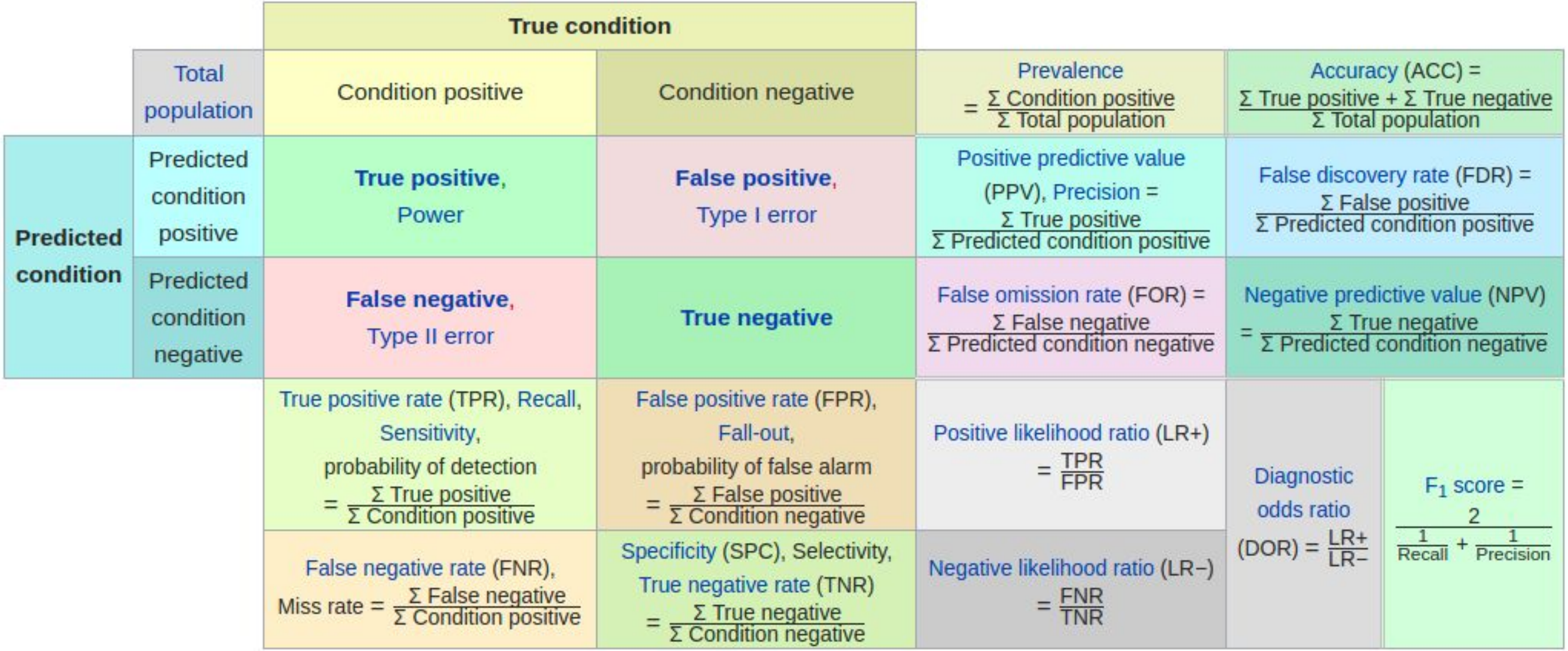
Next time

Results



- **SVM(ngram)** - Support Vector Machines classifier using text features including unigrams and bigrams
- **SVM(BF)** - Using many behavioral features including the number of reviews per day, rating deviation, content similarity, etc.
- **SVM(ngram+BF)** - combined behavioral features with ngram text features to improve the results.
- **PU-LEA** - The first Positive-Unlabeled learning model applied in review spam detection is PU-LEA.
- **LHMM (UT)** - Using the uniform transition (UT) probability in LHMM rather than that learned from data
- **LHMM** - The proposed LHMM model, Transition probabilities are learned from the training data
- **LHMM (MG)** - Just as LHMM, but the observed variables are co-bursting signals
- **CHMM** - This is the Coupled HMM model

Confusion matrix



		True condition				
	Total population	Condition positive	Condition negative	Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$	Accuracy (ACC) $= \frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$	
Predicted condition	Predicted condition positive	**True positive**, Power	**False positive**, Type I error	Positive predictive value (PPV), Precision $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$	False discovery rate (FDR) $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$	
	Predicted condition negative	**False negative**, Type II error	**True negative**	False omission rate (FOR) $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$	Negative predictive value (NPV) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$	
		True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$	False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$	Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$	Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR}-}$	F_1 score $= \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}}$
		False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$	Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$	Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$		

Are those real results?

Can I reproduce this method?