Social Media Anomaly Detection: Challenges and Solutions

Yan Liu ¹  Sanjay Chawla ²

¹Computer Science Department
Viterbi School of Engineering
University of Southern California

²Qatar Computing Research Institute and
School of Information Technologies
The University of Sydney

February 5, 2017
Acknowledgment

**Funding agencies:** NSF IIS-1134990 and IIS-1254206, DARPA, Australian Research Council, CapialMarkets CRC

**Special thanks to:**

Rose Yu
Aram Galstyan
Hanghang Tong
Leman Akoglu
Ching-Yung Lin
Zhen Wen
Zhen Wen
http://www-bcf.usc.edu/~liu32/SMAD.htm
Outline

1. Lecture 1: Introduction to social media anomaly detection
   - Overview of anomaly detection
   - Types and properties of social media data
   - Anomaly detection in network data
   - Anomaly detection in temporal data

2. Lecture 2: Recent advances in social media anomaly detection
What is Anomaly Detection?

Anomaly detection (or outlier detection)
Textbook definition: the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset.

Nice examples:
Generic Algorithm for Anomaly Detection

- Given a data set $D$, propose a model $M(D)$ which "generates" the data.
- Thus if $o \in D$ then let $\hat{o}$ be prediction from $M(D)$.
- $o$ is anomalous if $\|o - \hat{o}\|$ is large.
- Challenges of anomaly detection: outliers often have disproportional impact on the estimation of $M(D)$. 

![Diagram illustrating anomaly detection](image-url)
Challenges in Anomaly Detection

The reality is:
You never know what you are looking for. Anomaly detection may be more of “an art” than “the science”.

Issues with Existing Approaches
Most existing approaches to anomaly detection suffer from a series of shortcomings:

- **Sensitiveness**: high false alarm rate
- **Interpretation**: statistical test results with very limited insights about the detected anomaly
- **Scalability**: challenging for high-dimensional streaming data
Tutorial Themes

1. Special properties of social media anomaly detection:
   - We will provide concrete examples of social media anomaly detection

2. State-of-art techniques in anomaly detection:
   - We will address the issues in existing approaches

3. Working systems and competitions:
   - We will share practical scenarios and lessons learned
Outline

1. Lecture 1: Introduction to social media anomaly detection
   - Overview of anomaly detection
   - Types and properties of social media data
     - Anomaly detection in network data
     - Anomaly detection in temporal data

2. Lecture 2: Recent advances in social media anomaly detection
Social Media Data Types

Large-scale social media data usually consist of three data types: *structured data, unstructured texts and networks* labeled (sometimes) with temporal or/and spatial tags.
Examples of Social Media Anomaly Detection

Example 1: Bot detection
Examples of Social Media Anomaly Detection

Example 2: Compromised account detection
Examples of Social Media Anomaly Detection

Example 3: Group Review Spamming
Examples of Social Media Anomaly Detection

Example 4: Organized Viral Campaign
Examples of Social Media Anomaly Detection

Example 5: Bullying on Social Media
Categorization of Social Media Anomaly Detection

Based on the anomaly type, we have
- Point anomaly detection
- Group anomaly detection

Based on the input format, we have
- Activity-based: assume individuals are marginally independent
- Graph-based: account for relational information represented by graphs

Based on the temporal factor, we have
- Static information: one snapshot of the social network
- Dynamic information: time series observations of the social network
Challenges in Social Media Anomaly Detection

In addition to the challenges of classical anomaly detection tasks, social media also lead to new challenges:

- Heterogeneous data with rich and complex information
- Beyond the typical iid assumptions
- Very limited labeled examples or benchmark datasets
- Varieties and dynamics in anomalies
Outline

1. Lecture 1: Introduction to social media anomaly detection
   - Overview of anomaly detection
   - Types and properties of social media data
   - Anomaly detection in network data
   - Anomaly detection in temporal data

2. Lecture 2: Recent advances in social media anomaly detection
Overview of Graph Anomaly Detection

Graph Anomaly Detection

Static graphs
- Plain
  - Feature based
    - Structural features
    - Recursive features
  - Community based
- Attributed
  - Structure based
    - Substructures
    - Subgraphs
  - Community based

Dynamic graphs
- Plain
  - Distance based
    - Feature-distance
    - Structure distance
  - Structure based
    - “phase transition”

Credits: Akoglu et al, ASONAM Tutorial
Static Plain Graph

Feature Based Anomaly:
- Oddball [Akoglu et al. (2010)]
- Recursive structural features [Henderson et al. (2011)]

Community Based Anomaly:
- Bipartite graphs: neighborhood formation [Sun et al. (2005)]
- Non-negative residual matrix factorization [Tong and Lin (2011)]
- Anti-social communications [Ding et al. (2012)]
Static Attributed Graph

Substructure and subgraphs
- Minimum Descriptive Length (MDL) [Noble and Cook (2003)]
- MDL and probabilistic measure [Eberle and Holder (2007)]

Community outliers
- Probabilistic models [Gao et al. (2010)]
- PICS: cohesive clusters [Akoglu et al. (2012)]
Dynamic Graph

Distance based

- Graph distance: weight distance etc [Noble and Cook (2003)]
- ARIMA model [Pincombe (2005)]
- Scan statistics [Park et al. (2008)]

Structure based

- Eigen-space-based events [Idé and Kashima (2004)]
- GraphScope: matrix factorization [Sun et al. (2007)]
Outline

1. Lecture 1: Introduction to social media anomaly detection
   - Overview of anomaly detection
   - Types and properties of social media data
   - Anomaly detection in network data
   - Anomaly detection in temporal data

2. Lecture 2: Recent advances in social media anomaly detection
Temporal Data Anomaly Detection

**Point anomaly detection**
- Markov process
  - Bayes one-step Markov [Schonlau et al. (2001)]
  - Hybrid multi-step Markov [Ju and Vardi (2001)]
- Poisson process [Ihler et al. (2006)]
- Compression [Schonlau et al. (2001)]
- Probabilistic suffix tree (PST) [Sun et al. (2006)]
- Temporal dependence [Qiu et al. (2012)]
Temporal Data Anomaly Detection

Group anomaly detection

- Scan statistics [Das et al. (2009); Friedland and Jensen (2007)]
- Density estimation
  - Multinomial genre model (MGM) [Xiong et al. (2011a)]
  - Flexible genre model (FGM) [Xiong et al. (2011b)]
  - Group Latent Anomaly Detection model (GLAD) [Rose et al. (2014)]
  - One class support measure machine (OCSMM) [Muandet and Schölkopf (2013)]
Outline

1. Lecture 1: Introduction to social media anomaly detection

2. Lecture 2: Recent advances in social media anomaly detection
   - Point anomaly detection in social media
   - Group anomaly detection in social media
   - Fake news detection
   - Applications and systems
Point Anomaly Detection

Definition

Point anomaly detection aims to detect suspicious individuals, whose behavioral patterns deviate significantly from the general public.

Eg 1: Unusual file access
Eg 2: Abnormal network communication
Outline of Point Anomaly Detection

Activity-based Point Anomaly

Graph-based Point Anomaly
- Static graph
- Dynamic graph
Activity-based Point Anomaly Detection

Statistical hypothesis testing framework:
- Markov process
  - Bayes one-step Markov [Schonlau et al. (2001)]
  - Hybrid multi-step Markov [Ju and Vardi (2001)]
- Poisson process [Ihler et al. (2006)]
- Compression [Schonlau et al. (2001)]
- Probabilistic suffix tree (PST) [Sun et al. (2006)]
- Temporal dependence [Qiu et al. (2012)]

Comments
The activity sequences of each user are modeled under Markov assumption, which may suffer from rapid explosion in the dimension of the parameter space.
Markov Process

Application in detecting masquerades from UNIX commands usage records.

Bayes one-step Markov

**Null hypothesis**: one-step Markov process, the command of a user at current time relates to his previous command

**Alternative hypothesis**: multinomial distribution with Dirichlet prior

**Testing statistics**: the Bayes factor

Hybrid multi-step Markov

**Null hypothesis**: hybrid Markov model

**Alternative hypothesis**: commands are generated from other users

**Testing statistics**: combined statistics of the hybrid Markov model
Probabilistic Suffix Tree (PST)

Application in detecting outliers from a set of alphabetical sequences

Concepts

Edge → symbol in the alphabet
Node → string
Node distribution → the conditional probability of seeing a symbol right after the string label
Granger Graphical Models

**Basic idea**: Graphical modeling using the notions of Granger causality and methods of variable selection

**Granger Causality**: Cause happens prior to its effects [Granger 1969, 1980]. A time series $y$ is the *Granger* Cause of another time series $x$ if the past values of $y$ are helpful in predicting the future values of $x$ given its own past.

Practically, we perform the following two auto-regressions:

$$x_t = \sum_{l=1}^{L} a_l x_{t-l}$$

$$x_t = \sum_{l=1}^{L} a'_l x_{t-l} + \sum_{l=1}^{L} b'_l y_{t-l},$$

If Eq. (2) is a significantly better model than Eq. (1) (by statistical significance test), we determine that time series $y$ Granger causes time series $x$. 
Lasso-Granger [Arnold et al, KDD 2007]: Given $P$ time series $x^{(1)}, \ldots, x^{(P)}$ of length $T$, we can determine the Granger relationships of $x^{(i)}$ by performing the penalized auto-regression as follows:

$$
\min_{\{a_i\}} \sum_{t=L+1}^{T} \left\| x_t^{(i)} - \sum_{j=1}^{P} \beta_{i,j}^\top x_{t,Lagged}^{(j)} \right\|^2 + \lambda \| \beta_i \|_1,
$$

where

$$
x_{t,Lagged}^{(j)} = \left[ x_{t-L}^{(j)}, \ldots, x_{t-1}^{(j)} \right].
$$

Major advantages

- Variable selection can be efficiently achieved for high-dimensional time series
- Consistency analysis [Arnold et al, KDD 2007; Bahadori and Liu, 2012]
  
  Lasso-Granger: $P[\text{Error}] = o(c'L \exp(-T^v))$ for some $0 \leq v < 1$.
  
  Significant test: $P[\text{Error}] = o(c' \sqrt{T - L} \exp(-c^2(T - L)/2))$

Learning is possible even when the dimension $P$ is significantly larger than $T$!
Granger Graphical Models for Anomaly Detection

- Use Granger-lasso on training data: learn the coefficient $\hat{\beta}_i^{(a)}$ for each variable $x_i$ using lasso regression;
- Use constrained regression on the test data to learn another sets of coefficients $\hat{\beta}_i^{(b)}$:
  - Neighborhood similarity ($\epsilon_0 << \epsilon_1$):
    $$\sum_{j \in I_0} |\beta_{i,j}^{(b)}| \leq \epsilon_0, \sum_{j \in I_1} |\beta_{i,j}^{(b)}| \leq \epsilon_1,$$
  - Coefficient similarity:
    $$\sum_j |\beta_{i,j}^{(a)} - \beta_{i,j}^{(b)}| \leq \epsilon,$$
- Anomaly score: KL-divergence
  $$d_{i}^{ab} \equiv \int dx_i \frac{p(a)(x_i|X_L^{lagged})}{p(b)(x_i|X_L^{lagged})} \ln \frac{p(a)(x_i|X_L^{lagged})}{p(b)(x_i|X_L^{lagged})}$$
- Threshold: estimate the score distribution of training data; use 95% quantile as a threshold
Outline of Point Anomaly Detection

Activity-based Point Anomaly

Graph-based Point Anomaly

- Static graph
- Dynamic graph
Static Graph-based Point Anomaly Detection

Represent the relational information by graphs:

- Power law [Akoglu and McGlohon (2009); Akoglu et al. (2010)]
- Random walk [Moonesinghe and Tan (2008); Sun et al. (2005)]
- Hyper-graph [Silva and Willett (2008b,a)]
- Spatial auto-correlation [Sun and Chawla (2004); Chawla and Sun (2006)]

Comments

Consider not only the activity of individual users but also their interactions. Relies on nodes’ feature engineering from the graph. Strong assumptions on the graph generating process.
Power Law

Application in detecting anomalous nodes in subgraphs

1. Investigates the number of nodes $N_i$, the total weight $W_i$ and number of edges $E_i$ of the egonet $G_i$.

2. Defines the normal neighborhoods patterns: e.g. the Egonet Density Power Law (EDPL) pattern for $N_i$ and $E_i$: $E_i \propto N_i^\alpha$, $1 \leq \alpha \leq 2$; the Egonet Weight Power Law (EWPL) pattern for $W_i$ and $E_i^\beta$, $\beta \geq 1$.

3. Takes the distance-to-fitting-line as a measure to score the nodes in the graph.

Comments

Fitting of power law and the calculation of anomaly score is computationally efficient, easily fail if the network does not obey the power law.
Hyper-Graph

Definition

A hypergraph is a generalization of a graph in which an edge can connect any number of vertices.
Hyper-Graph

Application in detecting anomalous meetings in very large social networks

- Define $g(x)$ as the probability mass function of the meetings evaluated at a hype-edge $x$
- Define the distribution of the meetings as a two-component mixture: $g(x) = (1 - \pi)f(x) + \pi \mu(x)$, with $f(x)$ as nominal distribution, $\mu(x)$ as the anomalous distribution, $\pi$ as the mixture parameter
- $\mu(x)$: uniform distribution, $f(x)$: nonparametric density estimator
- Learn the likelihood of each observation using variational EM algorithm
- Anomalous score: model likelihood

Comments

A concise representation of complex interactions among multiple nodes, only applies to binary relationships where an edge is either present or missing.
Spatial Auto-correlation

Application in detecting spatial outliers, e.g. local anomalous counties from census data

1. Spatial neighborhood resembles the neighborhood defined in graph
2. Spatial Local Outlier Measure (SLOM): “stretched” distance between the point and its neighbors $\tilde{d}(a)$ and oscillating parameters $\beta(o)$
3. Use SLOM as anomalousness score to detect spatial outliers

Comments

SLOM captures the spatial autocorrelation and spatial heteroscedasticity (non-constant variance). Local spatial statistics would suffer from the “curse of dimensionality”.
Outline of Point Anomaly Detection

Activity-based Point Anomaly

Graph-based Point Anomaly
- Static graph
- Dynamic graph
Dynamic Graph-based Point Anomaly Detection

Three main categories [Bilgin and Yener (2010)]:

- Time series analysis of graph data
  - ARIMA process (Pincombe, 2005)
  - graph eigenvectors (Idé and Kashima, 2004)
- GraphScope: Minimum description length (MDL) (Sun et al., 2007)
- Window based approaches: scan statistics (Park et al., 2008)
Time Series Analysis

ARMA process (Pincombe, 2005)

1. Constructs a time series of changes for each graph topology distance measures.
2. Modeled each time series with an ARMA process.
3. Set up a residual threshold for the goodness of model fitting for time series.

Graph eigenvector (Idé and Kashima, 2004)

1. Define a time evolving dependency matrix from graphs.
2. Extract the principal eigenvector $u(t)$ as the “activity” vector.
3. Define the typical pattern as a linear combination of the past activity vectors.
4. Calculates the dissimilarity of the present activity vector from this typical pattern as anomalous score.
GraphScope: Minimum Description Length

Application in detecting the change points in a stream of graph series.

Concepts

**Graph segment**: One or more graph snapshots;

**Change point measure**: the encoding cost for $G^{(s)} \cup \{G^{(t)}\}$ as $c_n$ and $G^{(t)}$ as $c$. If $c_n - c_o < c$, the new graph is included in the current segment.

Rationale

Whether it is easier to include a new graph into the current graph segment or to start a new graph segment. If a new graph segment is created, it is treated as a change point.
Minimum Description Length

1. Compute the encoding cost of including a new graph into the current graph segment
2. Compute the encoding cost of starting a new graph segment
3. Compare the two costs and flag change point

Figure 2: Notation illustration: A graph stream with 3 graphs in 2 segments. First graph segment consisting of $G^{(1)}$ and ... terms are retained and this is the optimal length, if the range of $x$ is unknown [19]
Window based approach

Scan statistics
Slide a small window over local regions, computing certain local statistic for each window. The supremum or maximum of these locality statistics is known as the scan statistic.

Scan region: closed $k$th-order neighborhood of vertex $v$ in graph $D = (V, E)$: $N_k[v; D] = \{w \in V(D) : d(v, w) \leq k\}$. where $d(v, w)$ is the minimum directed path length from $v$ to $w$ in $D$.

Locality statistics: any digraph invariant $\Psi_k(v)$ of the scan region. For instance, the out degree of the digraph can be one such invariant locality statistics.

Comments
An intuitively appealing method to evaluate dynamic graph patterns, need to pre-specify a window width before one looks at the data.
Outline

1. Lecture 1: Introduction to social media anomaly detection

2. Lecture 2: Recent advances in social media anomaly detection
   - Point anomaly detection in social media
   - Group anomaly detection in social media
   - Fake news detection
   - Applications and systems
Group Anomaly Detection

Definition

Group anomaly or "collective anomaly" detection in social network aims to discover groups of participants that collectively behave anomalously. (Chandola et al. 2007).

The problem is challenging because:

- We do not know beforehand any members of a malicious group;
- The members of anomalous groups may change over time;
- Usually no anomaly can be detected when we examine individual member.
Activity-based Group Anomaly Detection

- Scan statistics [Das et al. (2009)]
- Density estimation
  - Multinomial genre model (MGM) [Xiong et al. (2011a)]
  - Flexible genre model (FGM) [Xiong et al. (2011b)]
  - Group Latent Anomaly Detection model (GLAD) Rose et al. (2014)
  - One class support measure machine (OCSMM) [Muandet and Schölkopf (2013)]
Density Estimation

**MGM**
Model groups as a mixture of Gaussian distributions with different mixture rates following the paradigm of latent models.

**FGM**
Extend MGM to with more flexibility in the generation of topic distributions.

**GLAD**
Infer the group membership and roles of each user automatically.

**OCSMM**
Generalize one-class support vector machine (OCSVM), compute the kernel of Gaussian distributions and apply SVM in a probability measure space.
Multinomial Genre Model (MGM)

Assumptions:

- **Groups are pre-computed**

---

**Algorithm 1 Generative process for MGMM**

```plaintext
for m = 1 to M do
    - Choose a group type \{1, \ldots, T\} \ni Y_m \sim \mathcal{M}(\pi)
    - Let the topic distribution \(\theta_m = \chi_{Y_m} \in \mathbb{S}^K\).
    - Choose \(N_m\), the number of points in the group \(G_m\). (\(N_m\) can be random, e.g. sampled from a Poisson distribution).
    for n = 1 to \(N_m\) do
        - Choose a galaxy type \(Z_{m,n} \in \{1, \ldots, K\}\), \(Z_{m,n} \sim \mathcal{M}(\theta_m)\).
        - Generate a galaxy feature \(X_{m,n} \in \mathbb{R}^f\), \(X_{m,n} \sim P(X_{m,n}|\beta, Z_{mn}) = \mathcal{N}(\beta_{Z_{m,n}}^\mu, \beta_{Z_{m,n}}^\Sigma)\).
    end for
end for
```
Flexible Genre Model (FGM)

Assumptions:
- Groups are *pre-computed*

Flexible Genre Model (FGM)

For each **group**:

1. Draw a genre
   \[ 1, 2, \ldots, T \ni y_m \sim \mathcal{M}(\pi) \]
2. Draw topic distribution for
   \[ y_m : S^K \ni \theta_m \sim Dir(\alpha_{y_m}) \]
3. Draw \( K \) topics
   \[ \{\beta_{mk} \sim P(\beta_{mk} | \nu_k)\}_{k=1, 2, \ldots, K} \]
4. For each **point** in group:
   1. Draw topic membership:
      \[ z_{mn} \sim \mathcal{M}(\theta_n) \]
   2. Generate point
      \[ x_{m,n} \in P(x_{m,n} | \beta_m, z_{m,n}) \]

Model Parameters
- \( \mathcal{M}(\pi) \) - Multinomial
- Each *genre* - Dirichlet
- Topic generators \( P(. | \nu) \)- Gaussian Inverse Wishart
- Point generators \( P(x_n | \beta_k) \) - Multivariate Gaussian
Flexible Genre Model (FGM)

Inference and Learning Parameters

- Approximate inference of latent variables (*Gibbs Sampling*)
- Use samples to learn parameters (*Single step Monte Carlo EM*)

Anomaly Detection

- Point based anomaly score:
  - Infer the topics ({$\beta_{m,k}$})
  - Compute negative log likelihood for all $\beta_{m,k}$ w.r.t. $\eta_k$
  - Rationale: If group contains anomalous points then corresponding topics will have low probability under $\eta$

- Distribution based anomaly score:
  - Infer the topic distribution $\theta_m$
  - Compute negative log likelihood w.r.t. $\alpha$
  - Rationale: An anomalous group will be unlikely to be generated from any genre
GLAD: Joint Models for Activity and Networks

Group latent anomaly detection model (GLAD) [Rose et al. (2014)]

**Concept of Role:**
1. Latent component in node features
2. Similar to an article topic

**Modeling Principal:** A group is modeled as a mixture of roles, with same of roles but different role mixture rate

**Definition of Group Anomaly**
Group anomaly has a *role mixture rate* pattern that does not conform to the majority of other groups.
Group Latent Anomaly Detection (GLAD0)

\[ \pi_p \propto \text{Dirichlet}(\alpha), \]
\[ G_p \propto \text{Multinomial}(\pi_p), \]
\[ R_p \propto \text{Categorical}(\theta_{G_p}), \]
\[ Z_{p\rightarrow q} \propto \text{Multinomial}(\pi_p), \]
\[ Z_{p\leftarrow q} \propto \text{Multinomial}(\pi_p), \]
\[ Y_{p,q} \propto \text{Bernoulli}(B_{Z_{p\rightarrow q},Z_{p\leftarrow q}}), \]
\[ X_p \propto \text{Multinomial}(\beta_{R_p}) \]

- High computational cost
- Loose connection of MMSB and LDA components via the shared group membership
Group Latent Anomaly Detection (GLAD)

A more computationally efficient model design

\[ \pi_p \propto \text{Dirichlet}(\alpha), \ G_p \propto \text{Multinomial}(\pi_p), \ R_p \propto \text{Categorical}(\theta_{G_p}), \ Y_{p,q} \propto \text{Bernouli}(B_{G_p,G_q}), \ X_p \propto \text{Multinomial}(\beta_{R_p}) \]
Dynamic extension of GLAD (d-GLAD)

Temporal evolution of the role mixture rate for each group is modeled as a series of multivariate Gaussian distributions: \( \theta_{tm}^t \propto \text{Gaussian}(\theta_{tm}^{t-1}, \sigma) \)
Procedure

Use BIC to decide \# of groups and \# of roles

Learn GLAD / d-GLAD to infer role mixture rates

Rank groups with respect to the anomaly score

Perform significant test and raise alarms

Calculate Anomaly Score

- GLAD: expected likelihood of role distribution
  \[ \text{AnomalyScore}_{\text{GLAD}} \propto \sum_{p \in G} E_q[p(R_p|\theta)] \]

- d-GLAD: change of role mixture rate over time
  \[ \text{AnomalyScore}_{\text{d-GLAD}} \propto \|\theta_{m}^{t-1} - \theta_{m}^{t}\|_2 \]
How do we determine outlier groups? Clearly Higher-Order Statistics are required. We will use Kernel Mean Embedding (KME) to form Higher-Order Statistics.
Smallest enclosing hypersphere problem

- Given a set of points \( S = \{x_1, x_2, \ldots, x_n\} \in \mathbb{R}^d \). Find the smallest hypersphere that encloses \( S \).

\[
\begin{align*}
\min_{R, c} & \quad R^2 \\
\text{subject to} & \quad \|x_i - c\|_2^2 \leq R^2 \quad \forall i = 1, \ldots n
\end{align*}
\]

- Standard Approach through Lagrangian multiplier

\[
L(c, R, \lambda) = R^2 + \sum_{i=1}^{n} \lambda_i [\|x_i - c\|^2 - R^2]
\]

- Optimizing \( L \) yields:

\[
\sum_{i=1}^{n} \lambda_i = 1 \text{ and } c = \sum_{i}^{n} \lambda_i x_i
\]
Working in Dual Space

- One can work entirely in the dual space.
- In fact, the Lagrangian can be expressed as

\[ L(c, R, \lambda) = \sum_{i=1}^{n} \lambda_i <x_i, x_i> - \sum_{i,j=1}^{n} \lambda_i \lambda_j <x_i, x_j> \]

- Or if we generalize to a positive-semidefinite kernel \( k \) then

\[ L(c, R, \lambda) = \sum_{i=1}^{n} \lambda_i k(x_i, x_i) - \sum_{i,j=1}^{n} \lambda_i \lambda_j k(x_i, x_j) \]

- Solve the dual optimization problem to estimate \( \lambda^* \).
Detecting Outliers

- To determine whether a new entity $x$ is an outlier with respect to the set $S$, test if

$$g(x) = \left\langle x, \sum_{i=1}^{n} \lambda_i x_i \right\rangle - R^2 > 0$$

i.e.,

$$g(x) = \langle x, x \rangle - 2 \sum_{i \in sv} \lambda_i \langle x, x_i \rangle + \sum_{i,j=1}^{n} \langle x_i, x_j \rangle - R^2 > 0$$

or with a kernel $k$

$$g(x) = k(x, x) - 2 \sum_{i \in sv} \lambda_i k(x, x_i) + \sum_{i,j=1}^{n} k(x_i, x_j) - R^2 > 0$$
Kernel Mean Embedding for Group Outlier Detection
[Muandet et. al.]

- Let $P$ be a group of points $\{x_1, \ldots, x_n\}$.
- Let $\phi$ be the kernel for $P$, i.e., all matrices of the form $\phi(x_i, x_j)$ are positive semidefinite (non-negative eigenvalues).
- The Hilbert Space associated with $\phi$ is the closed linear space of $\{\phi(., x)|x \in \mathbb{R}^d\}$. This is known as the reproducing kernel hilbert space (RKHS).
- The distribution can be represented via the kernel mean in RKHS: $\frac{1}{n} \sum_{i=1}^{n} \phi(., x_i)$.
- For certain $\phi$ (Gaussian kernel), the mapping is injective one-to-one.
- Let $P_1 = \{x_1, \ldots, x_{n_1}\}$ and $P_2 = \{y_1, \ldots, y_{n_2}\}$ are two groups of size $n_1$ and $n_2$ then form a dot product between the two groups as
  $$\frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi(x_i, y_j)$$
Static Graph-based Group Anomaly Detection

Graph-based group anomaly detection techniques seek to jointly utilize these observations and detect anomalous groups in a unified framework.

- Minimum description length (MDL) [Chakrabarti (2004); Lin and Chalupsky (2003); Rattigan and Jensen (2005)]
- Anomalous substructure [Noble and Cook (2003); Eberle and Holder (2007)]
- Tensor decomposition [Maruhashi et al. (2011)]
Anomalous Substructure

Given a labeled graph, each node as a label identifying its type:

1. Start with a list holding 1-vertex substructures for each unique vertex label.
2. Modify the list by generating, extending, deleting or inserting vertices and edges.
3. Count the number of occurrences for substructures.
4. Define a score for a substructure $S$ in a graph $G$ as $F_2 = \text{Size}(S) \cdot \text{Occurrences}(S, G)$, which is simply the product of the total number of nodes within a substructure and its occurrences.
Tensor Decomposition

Given an M-mode tensor $\mathcal{X}$ of size $I_1 \times I_2 \times \cdots \times I_M$,

1. Performs CP decomposition of the tensor of rank $R$ as $\mathcal{X} \approx \sum_{r=1}^{R} \lambda_r (a_r(1) \times \cdots a_r(M))$, where $\{a_r(i)\}$ are rank-1 eigenvectors.

2. Transform the eigenscore vector plot (absolute value of eigenscore vs. attribute index) into the eigenscore histogram (absolute value of eigenscore vs. frequency count).

3. Conduct spike detection on the histogram.

Comment

Capture the complex structure in heterogeneous networks. But tensor decomposition problem itself can be NP-hard to solve.
Evolving networks can also provide insights into the temporal changes of groups. Detecting anomalously groups in dynamic graphs is more challenging, as the group structures are not fixed and the unusual patterns in the group can also change.

- Bipartite graph [Friedland and Jensen (2007); Liu et al. (2008)]
- t-partite graph [Xu et al. (2007); Kim and Han (2009)]
- Counting process [Heard et al. (2010)]
Bipartite graph

Application in finding corporate tribes

Given bipartite graph $G = (R \cup A, E)$, $R = \{r_i\}$: the entity representatives, $A = \{a_j\}$: attributes, $E$: edges with time annotation.

1. List the co-worker relationships in the graph for every pair $f_{ij} = (r_i, r_j)$

2. Create a new graph $H = (R, F)$, where $F = \{f_{ij}\}$ is annotated with individuals attribute and history information.

3. Define a significance score for each edge, which measures the significance or the anomalousness of shared jobs.

4. Identify significant edges and computing the significance score $c$ for each of them.

5. Pick a threshold $d$ for the scores and prune all the edges $f_{ij}$ for $c_{ij} < d$.

6. Flag the connected components in the remaining graph as anomalous groups.
Outline

1. Lecture 1: Introduction to social media anomaly detection

2. Lecture 2: Recent advances in social media anomaly detection
   - Point anomaly detection in social media
   - Group anomaly detection in social media
   - Fake news detection
   - Applications and systems
Fake news

- Farmer shoots 23-lb. grasshopper!
- Giant bug is 4 feet long!
- Fed-up fatties kill aerobics instructor!

- Amazing aquagirl catches fish with her teeth!
- Sizzling photos inside!
- Man fries eggs on his bald head!
- Heart of gold! Pro wrestling star Norman the Lunatic brings joy to a sick little girl

- Mom shoots herself to give daughter a new heart
- America's biggest horoscope!
Fake news is interesting

- Misinformation can affect public opinion
  - German government: "We are dealing with a phenomenon of a dimension that we have not seen before"
- Bots pollute with fake activity
- Normal people also participate
  - NYT reported on a college graduate who started writing fake stories for fun and calculated that he earned "about 1,000 an hour in web advertising revenue" 1

Fake news is challenging

Curators are often sophisticated:
- Maintained by real people
- Distributed among many sources
- Buy users to give (fake) promotion

Further,
- Definition is not clear
- No clear tell-tale signs

Majority are confident in their ability to recognize fake news

PEW RESEARCH CENTER
What is fake news?

The "right" definition of fake news is not clear.

1. **Story that is not true**
   - Urban legends, satire, bad reporting (journalistic mistakes)
   - Fully false or contains false statements?
   - e.g. The Onion

2. **An opinion expressed for financial gain**
   - Propaganda, click-bait
   - Can be gibberish or related to true events
   - e.g. Chinese government has been cited for buying 'fake' supporters

3. **A biased story**
   - Reporting of personal opinion of a news story

4. **Opposing viewpoint**

5. **A story that is malicious and not true**

Some have tried to distinguish using "false" vs. "fake" vs. "falsehood" vs. "rumor", and so on...
What is fake news?

[Rubin et al. (2015)] proposed a classification into three types:

1. **Serious fabrication**: tabloids, click-bait
2. **Large-scale hoax**: deceptive, malicious
3. **Humorous fakes**: satire

Historically existing work has focused on (1), but now there is renewed interest in (2).
Existing Approaches

Existing approaches are most naturally group by the information used.

Text

Graph

Activity

http://www.businessinsider.com/google-algorithm-change-fake-news-rankbrain-2016-12
https://medium.com/@d1gi/the-election2016-micro-propaganda-machine-383449cc1fba#.x7qo60x0x
The Bursty Dynamics of the Twitter Information Network, Myers et al
Text-based

These methods utilize linguistic properties to try to detect fake news. Extract some textual features and apply your favorite classifier.

- **Stance detection** [Ferreira and Vlachos (2016)]
  - Detect a mismatch in between the headline and body text
  - for, against, observing
  - Logistic regression

- **Credibility ranking of tweets** [Gupta et al. (2014)]
  - Number of words, URLs, hashtags, emojis
  - Presence of swear words, pronouns
  - Use SVM-Rank with features.

<table>
<thead>
<tr>
<th>Linguistic features</th>
<th>example features</th>
</tr>
</thead>
<tbody>
<tr>
<td>posemo</td>
<td>love, nice, sweet</td>
</tr>
<tr>
<td>negate</td>
<td>no, not never</td>
</tr>
<tr>
<td>social</td>
<td>mate, talk, they, child</td>
</tr>
<tr>
<td>cogmech</td>
<td>cause, know, ought</td>
</tr>
<tr>
<td>excl</td>
<td>but, without, exclude</td>
</tr>
<tr>
<td>insight</td>
<td>think, know, consider</td>
</tr>
<tr>
<td>tentat</td>
<td>may be, perhaps, guess</td>
</tr>
<tr>
<td>see</td>
<td>view, saw, seen</td>
</tr>
<tr>
<td>hear</td>
<td>listen, hearing</td>
</tr>
</tbody>
</table>
Graph-based

The assumption is that fake news or users have a different connectivity than normal users.

- How fast does a rumor spread over a graph [Friggeri et al. (2014)]
- Which nodes/edges help fake news propagate [Karsai et al. (2013)]
- Fake news have different structural connectivity [Giasemidis et al. (2016)]
  - Triangles
  - Favoritism (retweeting the same set of users)

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_g$</td>
<td>Number of Nodes in the friendship network</td>
</tr>
<tr>
<td>$E_g$</td>
<td>Number of Links in the friendship network</td>
</tr>
<tr>
<td>$D_g$</td>
<td>Density of the friendship network</td>
</tr>
<tr>
<td>$C_g$</td>
<td>Clustering Coefficient of the friendship network</td>
</tr>
<tr>
<td>$I_g$</td>
<td>Median in-degree of the friendship network</td>
</tr>
<tr>
<td>$O_g$</td>
<td>Median out-degree of the friendship network</td>
</tr>
<tr>
<td>$F_l$</td>
<td>Fraction of nodes in the LCC</td>
</tr>
<tr>
<td>$V_l$</td>
<td>Number of nodes in the LCC</td>
</tr>
<tr>
<td>$E_l$</td>
<td>Number of links in the LCC</td>
</tr>
<tr>
<td>$D_l$</td>
<td>Density of nodes in the LCC</td>
</tr>
<tr>
<td>$C_l$</td>
<td>Clustering Coefficient in the LCC</td>
</tr>
<tr>
<td>$I_l$</td>
<td>Median in-degree in the LCC</td>
</tr>
<tr>
<td>$O_l$</td>
<td>Median out-degree in the LCC</td>
</tr>
<tr>
<td>$S_d$</td>
<td>Fraction of singletons in the diffusion network</td>
</tr>
<tr>
<td>$F_d$</td>
<td>Fraction of diffusion from low- to high-degree nodes</td>
</tr>
</tbody>
</table>
Activity-based

The information extracted captures the amount of activity occurring throughout time, for example, the number of retweets.

- Poisson process
  - Measure the number of retweets/shares over time [Bessi (2017)]
- Cluster based on activity
  - Colluding users will interact with similar items are similar times [Cao et al. (2014)]

<table>
<thead>
<tr>
<th>Symbols</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$N$</td>
<td>Total population of available users</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Probability of infection</td>
</tr>
<tr>
<td>$n_b$</td>
<td>Starting time of breaking news</td>
</tr>
<tr>
<td>$S_c$</td>
<td>Strength of external shock at birth (time $n_b$)</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>Background noise</td>
</tr>
<tr>
<td>$p_a$</td>
<td>Strength of interaction periodicity</td>
</tr>
<tr>
<td>$p_s$</td>
<td>Interaction periodicity offset</td>
</tr>
<tr>
<td>$q_a$</td>
<td>Strength of external shock</td>
</tr>
<tr>
<td>$q_p$</td>
<td>Periodicity of external shock</td>
</tr>
<tr>
<td>$q_s$</td>
<td>External shock periodicity offset</td>
</tr>
</tbody>
</table>
Mixture

These approaches combined structural, textual, temporal features.

- Apply feature selection with classification/clustering [Kwon et al. (2017), Giasemidis et al. (2016)]
- Feed into (recurrent) neural network [Ma et al. (2016)]
- Identify areas of connectivity with textually conflicting viewpoints [Jin et al. 2016]
We are just beginning

Fake news detection, particularly in the political context, is open and interesting...

- Microsoft sponsoring a panel “CONVERSATIONS: Proposition: We Can Solve The Fake News Problem”
- Fake news challenge (http://www.fakenewschallenge.org/)

Most of the work is focused on post-facto approaches for fake news identification, what about prediction and prevention?
Outline

1 Lecture 1: Introduction to social media anomaly detection

2 Lecture 2: Recent advances in social media anomaly detection
   • Point anomaly detection in social media
   • Group anomaly detection in social media
   • Fake news detection
   • Applications and systems
Example 1: Detecting Bots on Twitter

Bot detection: simple examples versus difficult examples
DARPA Bot Detection Challenge

Purpose

- Provide a high fidelity, simulated environment to evaluate the effectiveness of their strategies for identifying actors in an automated influence operation on Twitter

Data

- Simulated real-time feed of Twitter data via API
- The data is pulled from an actual influence challenge that took place in December 2014 and January 2015

Evaluation

- Accuracy and speed of identifying all the social bots in the dataset
PacSocial Influence Challenge Design

Two teams created and launched bots during the 4-week challenge. Teams were permitted to:

- A number of freedoms in order to authentically simulate an actual influence operation.
- Run any amount of bots to inhibit the spread of anti-vaccine content through the Twitter network.
- Update and change the behavior of bots during the course of the competition.
Data Description

- User information and the tweets: Approximately 7K users including bots and target network users
- Follower/friendship relationship: 4 weekly sequential series of snapshots of the network topology
Scoring: Accuracy and Speed

**Accuracy**

+1pt for every hit, -0.25pt for a false positive

**Speed**

Once a team identifies all the bots in the network, the team will be awarded +1 point for each day remaining in the competition.

Example: Team X finding all the bots five days before the end of the competition receives +5 points.

**Other requirement**

No limit on the number of guesses
Teams are ranked on their aggregate net points
Performance

Timeline:

![Timeline Diagram]

Results:

Accuracy Ranking
1. USC: 39.0 Points (39 hits, 0 misses)
2. [Other entries]
3. [Other entries]
4. [Other entries]
5. [Other entries]
6. [Other entries]

Speed Ranking
1. USC: 6 Points
2. [Other entries]
3. [Other entries]
4. [Other entries]

Final Ranking
1. USC: 45.0 Pts
2. [Other entries]
3. [Other entries]
4. [Other entries]
5. [Other entries]
6. [Other entries]

Contact: Aram Galstyan (USC/ISI)
Temporal features/statistics

- Inter-tweet time distribution for users
- Entropy based methods
- Reaction time for retweets/mentions
- Temporal anomalies in retweeting behavior
- Transfer entropy methods with tweet times

Follower/mention/retweet graph

- Calculate node centrality (Pagerank, etc)
- Analyze reciprocity relationships between friends/followers
- Analyze correlation between node centrality and activity measures
Combined text/network analysis

- Decompose #hashtag/user matrix to find topics/user groups
- LDA and other topic models
- Content Transfer

Sentiment analysis

- Classify tweet sentiment as pro vs. anti-vaccination
- Use unsupervised methods based on dictionaries
- Supervised by manually labeling some of the tweets
- Classify user sentiment as pro vs. anti-vaccination
Cluster-based Outlier Detection

Compute a list of simple features (22 total), such as

- Main API source
- Average tweeting activity (number of tweets per day)
- Number of mentioned users / number of tweets
- Ratio of mentioned tweets / retweets

Perform cluster-based outlier detection

- Conduct the outlier-resistant clustering via NMF
- Outliers that are difficult to assign to any cluster
### Aggregation

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
</tr>
</thead>
<tbody>
<tr>
<td>A55</td>
<td>Rumen</td>
<td>95</td>
<td>low</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
</tr>
<tr>
<td>A55</td>
<td>Rumen</td>
<td>95</td>
<td>low</td>
<td>95</td>
<td>95</td>
<td>95</td>
<td>95</td>
</tr>
</tbody>
</table>

#### Applications and systems
Lessons Learned

- Ensemble learning for unsupervised problems is challenging: How to best aggregate results from various methods?
- Current influence bots are, well, dumb with very limited NLP capabilities: Human-orchestrated campaigns are a more serious concern
Example 2: IBM ADAMS System

Architecture:

Contact: Ching-yung Lin (IBM Research)
Feature Extraction

<table>
<thead>
<tr>
<th>Category</th>
<th>Level</th>
<th>Examples</th>
<th>How-To</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structural</td>
<td>Local</td>
<td>degree, edge, weight</td>
<td>Ego-net [Oddball 2010]</td>
</tr>
<tr>
<td></td>
<td>Sub-graph</td>
<td>community, role</td>
<td>Matrix factorization, partition</td>
</tr>
<tr>
<td></td>
<td>Global</td>
<td>PageRank, centrality,</td>
<td>(Generalized) matrix-vector mul. [GBase 2011]</td>
</tr>
<tr>
<td>Content</td>
<td>Low-level</td>
<td>word frequency, tf/idf</td>
<td>Straight-forward</td>
</tr>
<tr>
<td></td>
<td>Topic-level</td>
<td>babble vs. commercial vs. research vs. social</td>
<td>SVD, LDA</td>
</tr>
<tr>
<td></td>
<td>High-level (semantic)</td>
<td>sentiment, event, usage,</td>
<td>Independent classifier, event modeling</td>
</tr>
</tbody>
</table>

- Whom does s/he talk to?
- What kind of roles does s/he play?
- What does s/he talk about?
- What is his/her opinion for a particular topic?
Learning Algorithm

Scenario 1: No labels
Density (LOF, LOCI)
Density Change (MALICE [He+ 2007])
Cluster-based algorithm

Scenario 2: One-class Labels
One-class SVM
LPU Learning [Liu+ 2003]

Scenario 3: Two-class Labels
Cost-sensitive learning [Chawla 2009]
Ensemble and Visualization

Ensemble:

Visualization:

- Y axis encodes topic significance
- X axis encodes time
- Height encodes the number of emails in the topic at this time
- A layer represents a topic
- Topic keywords
Summary

- Social media anomaly detection is an important and challenging task
- There are many existing work in related areas but the unique properties also raise new challenges
- Emerging topics
  - Bot detection
  - Compromised account detection
  - Yelp fake reviews
  - Uber fake ride


