The Ties that Bind
Characterizing Classes by Attributes and Social Ties

WWW
April, 2017

Aria Rezaei, Bryan Perozzi*, Leman Akoglu
Stony Brook University

*Now at Google.
Introduction
Our problem: *Characterizing Community Differences*

Proposed Method

Experimental Results

How does it relate to classification?
Attributed Graphs

- Nodes have a vector of properties
- Properties can be:
  - Scalar
    - e.g.: Product Rating, Salary, “#Publications in WWW”
  - Categorical
    - e.g.: Item Category, Ethnicity, “Ever Published in WWW?”
    - Can be changed into binary with more attributes!
Attributed Graphs Representation

- A matrix $A$ where:
  - $A_{vu}$ is 1 when edge $v \rightarrow u$ exists.
  - Entries can be:
    - $0/1 \rightarrow Unweighted Graphs$
    - Scalars $\rightarrow Weighted Graphs$

- A matrix $X$ where:
  - Row $X_v$ are attributes of node $v$
  - Entries can be:
    - $0/1 \rightarrow Binary Attributes$
    - Scalars $\rightarrow Scalar Attributes$
Characterizing Class Differences

- Given **subgraphs** of two **classes**
- Discover **attributes** that separate the classes
- These attributes are:
  - Mutually **exclusive**
  - Focused by subgraphs in each class
- Subgraphs can be:
  - **Ego-networks** of nodes in a class
  - **Local communities** around nodes in a class
  - **Pre-defined Communities**
Focus Attributes

- Nodes in a subgraph share a subset of attributes.
- We call this subset the **focus attributes**.
- We believe subgraphs from different classes exhibit *different* focus attributes.

![Diagram showing two communities: Community 1 with Crosswords, Tea parties, Gardening and Community 2 with Selfies, Partying, Video Games.](image-url)
Proposed Method
Method Outline

- There are four phases in our algorithm.
  1. Selecting subgraphs for two classes
  2. Identifying focus attributes in subgraphs.
  3. Separating attributes between the classes.
  4. Ranking the separated attributes in classes.

- Technically, subgraph selection is not the focus of our work.
Phase 1 – Subgraph Selection

- Identify nodes of each class
- Find communities around those nodes
  - Using local communities [Andersen+, 06]
  - Or ego-networks

\[ \text{G} \]

\[ g_{A1} \quad g_{A2} \]

\[ g_{B1} \quad g_{B2} \]
Phase 2 – Focus Attributes

- Extract focus attributes for each subgraph
- Rank attributes from *highest* focus to *lowest*
  - Using a ranking metric

Feature Extraction

Chess Biking

Rank

1. Chess
2. Biking
3. ...
Phase 2 – Focus Attributes (Cont’d)

- We use **Normality** weight vector [Perozzi+, 16]
  
  Idea: Subgraph **quality** is:
  
  - Internal
    - Structural **Density**
    - Attribute **Coherence**
  - External
    - Structural **Sparsity**
    - Attribute **Separation** from outside

- An attribute is a **focus** attribute when
  - It makes the subgraph **coherent**
  - It **separates** the subgraph from outside
Normality

- Normality’s formula

\[
N = I + E
\]

\[
I = \sum_{i \in C, j \in C} \left( A_{ij} - \frac{k_i k_j}{2m} \right) s(x_i, x_j | w)
\]

Internal Consistency

All node pairs in subgraph \(C\)

Null Model

Similarity Function, e.g:
- Dot product
- Kronecker’s
Normality

- Normality's formula

\[ N = I + E \]

\[ E = \sum_{i \in C, j \in C} \left( 1 - \min \left( 1, \frac{k_i k_b}{2m} \right) \right) s(x_i, x_b | w) \]

External Separability

Links from inside \( C \) to outside
Normality

- Normality’s formula

\[ N = \mathbf{w}_c^T \cdot (\hat{x}_I + \hat{x}_X) \]

- \[ \max \mathbf{N} \quad \text{s.t.} \quad \|\mathbf{w}_c\|_p = 1, \mathbf{w}_c(a) \geq 0, \forall a = 1, \ldots, d \]

  - **L1 norm** \[ \mathbf{w}_c(a) = 1, \text{one attribute with largest } \mathbf{x} \]
  - **L2 norm** \[ \mathbf{w}_c(a) = \frac{x(a)}{\sqrt{\sum_{x(i) > 0} x(i)^2}}, \text{all attributes with positive } \mathbf{x} \]
Phase 3 – Splitting Attributes

- Putting together vectors of all subgraphs
- We have a matrix of attribute weights for each class
Phase 3 – Splitting Attributes (Cont’d)

Idea:
- We don’t want attributes that are:
  - Relevant or irrelevant to both classes

Highly relevant to both. Not distinguishing.

Irrelevant to both. Not Interesting.
Phase 3 – Splitting Attributes (Cont’d)

Idea:

- we want attributes that are:
  - Relevant to one class and irrelevant to the other.
Objective Function

- Given a subset of attributes $S$
- Quality of a subgraph $g$ is:

$$N(g \mid S) = \sqrt{\sum_{a \in S} x(a)^2} = \|x[S]\|_2$$

2-norm of $x$ induced on the attribute subspace
attribute weight vector of $g$
Objective Function (Cont’d)

The quality of a separation will be:

$$\max_{A^+ \subseteq A, A^- \subseteq A} \frac{1}{p} \sum_{i \in S^+} \|x_i[A^+]\|_2 + \frac{1}{n} \sum_{j \in S^-} \|x_j[A^-]\|_2$$

Such that $A^+ \cap A^- = \emptyset$

$p = \text{number of subgraphs in class } +$

$n = \text{number of subgraphs in class } -$
Submodular Welfare Problem (SWP)

- **Definition:**

Given \(d\) items and \(m\) players having a **monotone** and **submodular** utility function \((w_i)\) over subsets of items. Partition the \(d\) items into \(m\) disjoint sets \((I_1, I_2, \ldots, I_m)\) in order to maximize:

\[
\sum_{i=1}^{m} w_i(I_i)
\]

- Our quality function \(N(g|S)\) is a **monotone** and **submodular** set function.
SWP History

- It’s known to be NP-hard.
- First approx. factor is $\frac{1}{2}$. [Lehman+, 2001]
- Improved to $(1 - 1/e)$ [Vondrak+, 08]
- No better results unless:
  - $P = NP$ [Khot+, 08]
  - Using exponentially many value queries [Mirrokni+, 08]
Faster Alternatives

1. Pre-normalized Weights
   - Simplify calculations by pre-computing attribute weights:
     - Instead of normalizing by a selected subset
     - Do it with all subgraphs of a class.
   - This makes the quality function modular
   - A linear sweep over attributes finds solution.
     - With time linear in $n$

\[ w_g(a) = \frac{x(a)}{\sqrt{\sum_{a \in A} x(a)^2}} \]
Faster Alternatives

1. Pre-normalized Weights
2. Top-k Attributes
   - In reality we might need only the top few attributes
   - We optimize the overall objective function
   - Adding attributes one by one
   - Until we have the top-k for two classes.
   - We use the lazy greedy method of [Leskovec+, 07]
Phase 4 – Ranking Attributes

- How **characterizing** is an attribute?
  - Compared to other attributes selected for a class

Relative Contribution Score (RC):

$$rc(a) = \frac{1}{p} \sum_{i \in S^+} x_i(a) - \frac{1}{n} \sum_{j \in S^-} x_j(a)$$

- Normalized contribution of $a$ to Class +
- Normalized contribution of $a$ to Class -
Experiments
Synthetic Experiments

- Three algorithms:
  1. SWA (Submodular Welfare Algorithm)
  2. Simplified (with pre-normalized weights)
  3. Top-k (only top-k attributes)

- Two Tests:
  - Optimality Test
    - How close are their results to the optimal results?
  - Computational Performance
    - How fast can they execute?
Optimality Test

We test the quality ratio to **optimal** result.

Both **SWA** and **Simplified** work near perfectly.

**Simplified** assigns all attr. to one class, ignoring **Diminishing Returns** property.
Computational Performance

- We test the performance of three methods
  - How their running time grows w.r.t $n$

![Graph showing computational performance](image)
Real-World Datasets

- Baseline Method
  - L1-Regularized Logistic Regression
  - Positive weights are given to class $+1$
  - Negative weights are given to class $-1$

- Datasets
  - Congress Co-sponsorship Network
  - Amazon Co-purchase Network
  - DBLP Co-authorship Network
Congress Co-sponsorship

- Bills in Congress
  - Each bill has a group of sponsors.
  - Each bill has a policy area tag.
- Nodes are congressmen
- Edges:
  - Two nodes connected if they co-sponsored a bill
- Attributes:
  - Party affiliation of congressmen
  - Policy areas of bills they sponsored.
Congress Co-sponsorship (Cont’d)

- 13 consecutive congress 2-year cycles.
- 435 congressmen
- 32 policy area tags, e.g:
  - National Security and Armed Forces
  - Environmental Protection
  - Foreign Affairs
  - …
Conservative vs. Liberal Ideas

Focus attributes reveal the contrast between Liberal and Conservative ideas

Democrats focus mostly on social programs

Republicans focus mostly on governance and finance
Focus of Political Parties

We can also see how issues lose/gain party focus

PARTY FOCUS ON ARMED FORCES

- War in Afghanistan
- War in Iraq
- Bombing of Iraq

1993 1995 1997 1999 2001 2003 2005 2007

Democrat
Republican
Amazon Co-purchase Network

- Nodes are *Amazon videos*
- Edges:
  - Between two products *co-purchased* together
- Attributes are product genres:
  - Content theme (Drama, Comedy, …)
  - Audience age range (10-12 Years, etc.)
  - …
Query: Animation vs. Classic

Ours

LASSO

Franchise names

Age range

Creator

Content
Query: Under 13 vs. Over 13

Attribute weight goes down as quality decreases

Ours

LASSO

Not much differentiation

Aria Rezaei

Characterizing Classes by Attributes and Social Ties
DBLP Co-authorship Network

- Nodes are **authors**.
- Links between **co-authors**.
- Attributes are **venues** an author publishes in:
  - WWW, KDD, …
Query: ICC vs. ICASSP Conferences

Conferences on:
- Networking
- Communications
- Mobile Tech.
- Sensors

ICASSP
- INTERSPEECH
- IEEE Trans. on Sig. Proc.
- LREC
- ICIP
- Sig. Proc.
- ICME
- EUSIPCO
- IGARSS
- ESANN
- DICTA

Conferences on:
- Speech Proc.
- Video Proc.
- Image Proc.
- Signal Proc.
- Linguistics

 ICC
Characterization vs. Classification
Characterization vs. Classification

- Regularized classifiers (e.g.: LASSO) can:
  - Find a sparse attribute subspace
  - With ranking weights
  - How is our work different?
Class Support and Confidence

Confidence → Prob. of belonging to class $c$ if $a$ is observed

$$Cfd(c, a) = Pr(c|a) = \frac{\#(c, a)}{\#(a)}$$

Support → Portion of nodes in class $c$ exhibiting $a$

$$Sup(c, a) = \frac{\#(c, a)}{\#(c)}$$
Class Support and Confidence

Class Confidence

**Relative Confidence**

\[
CC(c^+, a) = \Pr(c^+ | a) - \Pr(c^- | a)
\]

Class Confidence

**Relative Support**

\[
CS(c^+, a) = \text{Sup}(c^+, a) - \text{Sup}(c^-, a)
\]

We believe classifiers focus on **confidence** while we focus on **support**.
Proposed Method vs. LR

Class Support

Class Confidence

Animation  Classics  Under 13  Over 13  ICASSP  ICC

Proposed
LR
Questions?

Thank you.

code and datasets
www.cs.stonybrook.edu/~arezaei