Analysing Online Social Network Data with Biclustering and Triclustering

Alexander Semenov\textsuperscript{1} \hspace{1cm} Dmitry Ignatov\textsuperscript{1}
Dmitry Gnatyshak\textsuperscript{1} \hspace{1cm} Jonas Poelmans\textsuperscript{2,1}

\textsuperscript{1}NRU Higher School of Economics, Russia
\textsuperscript{2}KU Leuven, Belgium
Motivation I

• There are large amount of network data that can be represented as bipartite and tripartite graphs
• Standard techniques like maximal bicliques search result in huge number of patterns (in the worst case exponential w.r.t. of input size)...
• Therefore we need some relaxation of this notion and good measures of interestingness of biclique communities
Motivation II

• Applied lattice theory provide us with a notion of formal concept which is the same thing as biclique
• Social Networks 18(3), 1996
  – V. Duquenne, Lattice analysis and the representation of handicap associations.
• Camille Roth et al., Towards Concise Representation for Taxonomies of Epistemic Communities, CLA 4th Intl Conf on Concept Lattices and their Applications, 2006
• And many other papers on application to social network analysis with FCA
Motivation III

• Concept-based bicluster (Ignatov et al., 2010) is a scalable approximation of a formal concept (biclique)
  – Less number of patterns to analyze
  – Less computational time (polynomial vs exp.)
  – Manual tuning of bicluster (community) density threshold
  – Tolerance to missing (object, attribute) pairs

• For analyzing three-way network data like folksonomies we proposed triclustering (Ignatov et al., 2011)
Definition 1. **Formal Context** is a triple \((G, M, I)\), where \(G\) is a set of (formal) objects, \(M\) is a set of (formal) attributes, and \(I \subseteq G \times M\) is the incidence relation which shows that object \(g \in G\) possesses an attribute \(m \in M\).

**Example**

<table>
<thead>
<tr>
<th></th>
<th>Car</th>
<th>House</th>
<th>Laptop</th>
<th>Bicycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kate</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Mike</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Alex</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>David</td>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>
Formal Concept Analysis

Definition 2. Derivation operators (defining Galois connection)

\[ A^l := \{ m \in M \mid glm \text{ for all } g \in A \} \] is the set of attributes common to all objects in \( A \)

\[ B^l := \{ g \in G \mid glm \text{ for all } m \in B \} \] is the set of objects that have all attributes from \( B \)

Example

<table>
<thead>
<tr>
<th></th>
<th>Car</th>
<th>House</th>
<th>Laptop</th>
<th>Bicycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kate</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Mike</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alex</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>David</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>

\( \{Kate, Mike\}^l = \{Car\} \)

\( \{Laptop\}^l = \{Mike, Alex, David\} \)

\( \{Car, House\}^l = \emptyset _G \)

\( \emptyset _G ^l = M \)
Formal Concept Analysis

**Definition 3.** \((A, B)\) is a formal concept of \((G, M, I)\) iff:

\[ A \subseteq G, \ B \subseteq M, \ A^I = B, \text{ and } B^I = A. \]

\(A\) is the **extent** and \(B\) is the **intent** of the concept \((A, B)\).

\(\mathcal{B}(G, M, I)\) is a set of all concepts of the context \((G, M, I)\).

**Example**

<table>
<thead>
<tr>
<th></th>
<th>Car</th>
<th>House</th>
<th>Laptop</th>
<th>Bicycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kate</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Mike</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alex</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>David</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- A pair \(\{\text{Kate, Mike}\}, \{\text{Car}\}\) is a formal concept.
- \(\{\text{Alex, David}\}, \{\text{Laptop}\}\) **doesn't form a formal concept,** because \(\{\text{Laptop}\}^I \neq \{\text{Alex, David}\}\).
- \(\{\text{Alex, David}\} \{\text{House, Laptop}\}\) **is a formal concept**
FCA and Graphs

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kate</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Mike</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Alex</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>David</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

Formal Context

Formal Concept (maximal rectangle)

Bipartite graph

Biclique
Definition 4. A formal concept \((A,B)\) is said to be more general than \((C,B)\), that is \((A,B) \geq (C,D)\) iff \(A \subseteq C\) (equivalently \(D \subseteq B\)).

The set of all concepts of the context \((G, M, I)\) ordered by relation \(\geq\) forms a complete lattice \(\mathcal{B}(G,M,I)\) called concept lattice (Galois lattice).

Example

<table>
<thead>
<tr>
<th></th>
<th>Car</th>
<th>House</th>
<th>Laptop</th>
<th>Bicycle</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kate</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Mike</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Alex</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>David</td>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

\(\{\{Alex, David, Mike\}, \{Laptop\}\}\)

is more general than concept

\(\{\{Alex, David\} \{House, Laptop\}\}\)
## Concept Lattice Diagram

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kate</td>
<td>x</td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Mike</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Alex</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>David</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

- **a - Car**
- **b - House**
- **c - Laptop**
- **d - Bicycle**
Definition 1 If \((g, m) \in I\), then \((m', g')\) is called an object-attribute or \(oa\)-bicluster with density \(\rho(m', g') = \frac{|I \cap (m' \times g')|}{|m'| \cdot |g'|}\).
Biclustering Example

Since $(House, David)$ is in the context

$(House^l, David^l) = (\{Alex, David\}, \{House, Laptop, Bicycle\})$

$\rho(House^l, David^l) = 5/6$
Biclustering properties

• Number of all biclusters for a context \((G, M, I)\) not greater than \(|I| \leq 2^{\min\{|G|, |M|\}}\) formal concepts. Usually \(|I| \ll 2^{\min\{|G|, |M|\}}\), especially for sparse contexts.

• Probably **dense biclusters** \((\rho(bicluster) \geq \rho_{min})\) are good representation of communities, because all users inside the extent of every dense bicluster have almost all interests from its intent.
Definition 1. Triadic Formal Context is a quadruple \((G, M, B, Y)\), where \(G\) is a set of (formal) objects, \(M\) is a set of (formal) attributes, \(B\) is a set of conditions, and \(Y \subseteq G \times M \times B\) is the incidence relation which shows that object \(g \in G\) possesses an attribute \(m \in M\) under condition.

Example. Folksonomy as triadic context \((U, T, R, Y)\), where

- \(U\) is a set of users
- \(T\) is a set of tags
- \(R\) is a set of resources
### Concept forming operators in triadic case

To define triclusters we propose **box operators**

---

**Table 1. Prime and double prime operators of 1-sets**

<table>
<thead>
<tr>
<th>Prime operators of 1-sets</th>
<th>Their double prime counterparts</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m' = { (g, b)</td>
<td>(g, m, b) \in Y }$</td>
</tr>
<tr>
<td>$g' = { (m, b)</td>
<td>(g, m, b) \in Y }$</td>
</tr>
<tr>
<td>$b' = { (g, m)</td>
<td>(g, m, b) \in Y }$</td>
</tr>
</tbody>
</table>

---

\[
g^\Box = \{ g_i | (g_i, b_i) \in m' \text{ or } (g_i, m_i) \in b' \}
\]

\[
m^\Box = \{ m_i | (m_i, b_i) \in g' \text{ or } (g_i, m_i) \in b' \}
\]

\[
b^\Box = \{ b_i | (g_i, b_i) \in m' \text{ or } (m_i, b_i) \in g' \}.
\]
Triclustering
[Ignatov et al., 2011]

Let $\mathbb{K} = (G, M, B, Y)$ be a triadic context. For a certain triple $(g, m, b) \in Y$, the triple $T = (g^\square, m^\square, b^\square)$ is called a tricluster.

The density of a certain tricluster $(A, B, C)$ of a triadic context $\mathbb{K} = (G, M, B, Y)$ is given by the fraction of all triples of $Y$ in the tricluster, that is $\rho(A, B, C) = \frac{|I \cap A \times B \times C|}{|A| |B| |C|}$.

Table 2. A toy example with Bibsonomy data for users $\{u_1, u_2, u_3\}$, resources $\{r_1, r_2, r_3\}$ and tags $\{t_1, t_2, t_3\}$

\[
\begin{array}{c|c|c|c}
\text{t}_1 & \text{t}_2 & \text{t}_3 \\
\hline
\text{u}_1 & \times & \times \\
\text{u}_2 & \times & \times & \times \\
\text{u}_3 & \times & \times & \times \\
\end{array}
\]

\[
\begin{array}{c|c|c|c}
\text{t}_1 & \text{t}_2 & \text{t}_3 \\
\hline
\text{u}_1 & \times & \times & \times \\
\text{u}_2 & \times & \times & \times \\
\text{u}_3 & \times & \times & \times \\
\end{array}
\]

$T = (\{u_1, u_2, u_3\}, \{t_1, t_2, t_3\}, \{r_1, r_2, r_3\})$ with $\rho = 0.89$

One dense tricluster $VS\ 3^3 = 27$ formal triconcepts
Pseudo Triclustering for Social Networks

Let $K_{UI} = (U, I, X \subseteq U \times I)$ be a formal context which describes what interest $i \in I$ a particular user $u \in U$ has. Similarly, let $K_{UG} = (U, G, Y \subseteq U \times G)$ be a formal context which indicates what group $g \in G$ user $u \in U$ belongs to.

We can find dense biclusters as $(users, interests)$ pairs in $K_{UI}$ using oabiclustering algorithm which is described in Ignatov et. al (2010). These biclusters will be exactly groups of users that have similar interests. In the same way we can find communities of users which belong to similar groups of Vkontakte social network as dense biclusters $(users, groups)$.

To this end we need to mine a (formal) tricontext $K_{UIG} = (U, I, G, Z \subseteq U \times I \times G)$, where $(u, i, g)$ is in $Z$ iff $(u, i) \in X$ and $(u, g) \in Y$. A particular tricluster has a form $T_k = (i^X \cap g^Y, u^X, u^Y)$ for every $(u, g, i) \in Z$ with $\frac{|i^X \cap g^Y|}{\min(|i^X|, |g^Y|)} \geq \Theta$, where $\Theta$ is a predefined threshold between 0 and 1.
Algorithm

Formal context \((U_{ui}, I, X)\) (minimal density \(p_{ui}\))

Formal context \((U_{ug}, G, Y)\) (minimal density \(p_{ug}\))

Biclusters’ extent similarity \(\mu \geq C\)

Set of biclusters \(Bicl_{UI}\)

Set of pseudo-triclusters (Possible constraint: \(\rho_{avg} \geq \rho_{min}\))

Set of biclusters \(Bicl_{UG}\)
Algorithm

Let $\text{Bicl}_{UI}$ be a set of user-interest biclusters and $\text{Bicl}_{UG}$ be a set of user-group biclusters.

For each $(U_{ui}, I) \in \text{Bicl}_{UI}$ and $(U_{ug}, G) \in \text{Bicl}_{UG}$ triple $(U_{ui} \cap U_{ug}, I, G)$ is added to triclusters’ set if $U_{ui} \cap U_{ug} \neq \emptyset$ and

$$\mu = \frac{|U_{ui} \cap U_{ug}|}{|U_{ui} \cup U_{ug}|} \geq C, \ 0 \leq C \leq 1.$$

Thus, $\mu$ is used as a measure of quality of these pseudo-triclusters.

Another measure is an average density of biclusters:

$$\frac{\rho[(U_{ui}, I)] + \rho[(U_{ug}, G)]}{2}.$$

Test setting: Intel Core i7-2600 system with 3.4 GHz and 8 GB RAM

Constraints for the formal contexts used: $\rho \geq 0.5$. 
Pseudo-triclustering algorithm was tested on the data of Vkontakte, Russian social networking site. Student of two major technical and two universities for humanities and sociology were considered:

<table>
<thead>
<tr>
<th></th>
<th>Bauman</th>
<th>MIPT</th>
<th>RSUH</th>
<th>RSSU</th>
</tr>
</thead>
<tbody>
<tr>
<td># users</td>
<td>18542</td>
<td>4786</td>
<td>10266</td>
<td>12281</td>
</tr>
<tr>
<td># interests</td>
<td>8118</td>
<td>2593</td>
<td>5892</td>
<td>3733</td>
</tr>
<tr>
<td># groups</td>
<td>153985</td>
<td>46312</td>
<td>95619</td>
<td>102046</td>
</tr>
</tbody>
</table>

Data
# Biclustering results

<table>
<thead>
<tr>
<th>ρ</th>
<th>Bauman</th>
<th>MIPT</th>
<th>RSUH</th>
<th>RSSU</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Time</td>
<td>#</td>
<td>Time</td>
<td>#</td>
</tr>
<tr>
<td>0.0</td>
<td>9188</td>
<td>8863</td>
<td>1874458</td>
<td>248077</td>
</tr>
<tr>
<td>0.1</td>
<td>8882</td>
<td>8331</td>
<td>1296056</td>
<td>173786</td>
</tr>
<tr>
<td>0.2</td>
<td>8497</td>
<td>6960</td>
<td>966000</td>
<td>120075</td>
</tr>
<tr>
<td>0.3</td>
<td>8006</td>
<td>5513</td>
<td>788008</td>
<td>85227</td>
</tr>
<tr>
<td>0.4</td>
<td>7700</td>
<td>4308</td>
<td>676733</td>
<td>59179</td>
</tr>
<tr>
<td>0.5</td>
<td>7536</td>
<td>3777</td>
<td>654047</td>
<td>53877</td>
</tr>
<tr>
<td>0.6</td>
<td>7324</td>
<td>2718</td>
<td>522110</td>
<td>18586</td>
</tr>
<tr>
<td>0.7</td>
<td>7250</td>
<td>2409</td>
<td>511711</td>
<td>15577</td>
</tr>
<tr>
<td>0.8</td>
<td>7217</td>
<td>2326</td>
<td>508368</td>
<td>14855</td>
</tr>
<tr>
<td>0.9</td>
<td>7246</td>
<td>2314</td>
<td>507983</td>
<td>14691</td>
</tr>
<tr>
<td>1.0</td>
<td>7236</td>
<td>2309</td>
<td>511466</td>
<td>14654</td>
</tr>
</tbody>
</table>
## Pseudo triclustering results

| \( \mu \) | **Bauman** | | **MIPT** | | **RSUH** | | **RSSU** |
|---|---|---|---|---|---|---|
|  | Time, ms | Count | Time, ms | Count | Time, ms | Count | Time, ms | Count |
| 0.0 | 3353426 | 230161 | 77562 | 24852 | 256801 | 35275 | 183595 | 55338 |
| 0.1 | 76758 | 10928 | 35137 | 5969 | 62736 | 5679 | 18725 | 5582 |
| 0.2 | 80647 | 8539 | 31231 | 4908 | 58695 | 5089 | 16466 | 3641 |
| 0.3 | 77956 | 6107 | 27859 | 3770 | 53789 | 3865 | 17448 | 2772 |
| 0.4 | 60929 | 31 | 2060 | 12 | 9890 | 14 | 13585 | 12 |
| 0.5 | 66709 | 24 | 2327 | 10 | 9353 | 14 | 12776 | 10 |
| 0.6 | 57803 | 22 | 2147 | 8 | 11352 | 14 | 12268 | 10 |
| 0.7 | 68361 | 18 | 2333 | 8 | 10778 | 12 | 13819 | 4 |
| 0.8 | 70948 | 18 | 2256 | 8 | 9489 | 12 | 13725 | 4 |
| 0.9 | 65527 | 18 | 1942 | 8 | 10769 | 12 | 11705 | 4 |
| 1.0 | 65991 | 18 | 1971 | 8 | 10763 | 12 | 13263 | 4 |
Pseudo triclustering results

Number of pseudo-triclusters for different values of $\mu$

- Bauman
- MIPT
- RSUH
- RSSU
Examples. Biclusters

• $\rho = 83,33\%$  
  Gen. pair: \{3609, home\}  
  G: \{3609, 4566\}  
  M: \{family, work, home\}

• $\rho = 83,33\%$  
  Gen. pair: \{30568, orthodox church\}  
  G: \{25092, 30568\}  
  M: \{music, monastery, orthodox church\}

• $\rho = 100\%$  
  Gen. pair: \{4220, beauty\}  
  G: \{1269, 4220, 5337, 20787\}  
  M: \{love, beauty\}
Examples. Tricluster

• Measures:
  \[ \mu : 100\%; \]
  Average \[ \rho : 54,92\% \]

Users: \{16313, 24835\}

Interests: \{sleeping, painting, walking, tattoo, hamster, impressions\}

Groups: \{365, 457, 624,\ldots, 17357688, 17365092\}
Conclusion

• It is possible to use pseudo-triclustering method for tagging groups by interests in social networking sites and finding tricommunities. E.g., if we have found a dense pseudo-trciluster (Users, Groups, Interests) we can mark Groups by user interests from Interests.

• It also make sense to use biclusters and tricluster for making recommendations. Missing pairs and triples seem to be good candidates to recommend potentially interesting users, groups and interests.
Conclusion

• The approach needs some improvements and fine tune in order to increase the scalability and quality of communities
  – Strategies for approximate density calculation
  – Choosing a good thresholds for n-clusters density and communities similarity
  – More sophisticated quality measures like recall and precision in Information Retrieval
• It needs comparison with other approaches like iceberg lattices (Stumme), stable concepts (Kuznetsov), fault-tolerant concepts (Boulicaut) and different n-clustering techniques from bioinformatics (Zaki, Mirkin, etc.)
• Current version also requires expert’s feedback on the output data analysis and interpretation
Questions?