Analysing Online Social Network Data with Biclustering and Triclustering

Alexander Semenov[1] Dmitry Ignatov[1]
Dmitry Gnatyshak[1] Jonas Poelmans[2,1]

[1]NRU Higher School of Economics, Russia
[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
- L. C. Freeman, D. R. White. **Using Galois Lattices to Represent Network Data** *Sociological Methodology 1993* (23).
- *Social Networks* 18(3), 1996
  - L. C. Freeman, **Cliques, Galois Lattices, and the Structure of Human Social Groups.**
  - V. Duquenne, **Lattice analysis and the representation of handicap associations.**
  - D. R. White. **Statistical entailments and the Galois lattice.**
- J.W. Mohr, Vincent D. **The duality of culture and practice: Poverty relief in New York City**, **1888—1917** *Theory and Society*, 1997
- 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)

Formal Concept Analysis

[Wille, 1982, Ganter & Wille, 1999]

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 $g \in G$ posseses an attribute $m \in M$.

Example

	Car	**House**	**Laptop**	**Bicycle**
Kate	x			x
Mike	x		x	
Alex		x	x	
David		x	x	x

Formal Concept Analysis

Definition 2. Derivation operators (defining Galois connection)

$A^I := \{ m \in M \mid gIm \text{ for all } g \in A \}$ is the set of attributes common to all objects in A

$B^I := \{ g \in G \mid gIm \text{ for all } m \in B \}$ is the set of objects that have all attributes from B

Example

	Car	**House**	**Laptop**	**Bicycle**
Kate	x			x
Mike	x		x	
Alex		x	x	
David		x	x	x

$\{Kate, Mike\}^I$ = {Car}

{Laptop}^{I} = $\{Mike, Alex, David\}$

$\{Car, House\}^I = \{\}_G$

$\{\}^I_G = M$

Formal Concept Analysis

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

$$A \subseteq G,\ B \subseteq M,\ A' = B, \text{ and } B' = A\ .$$

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

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

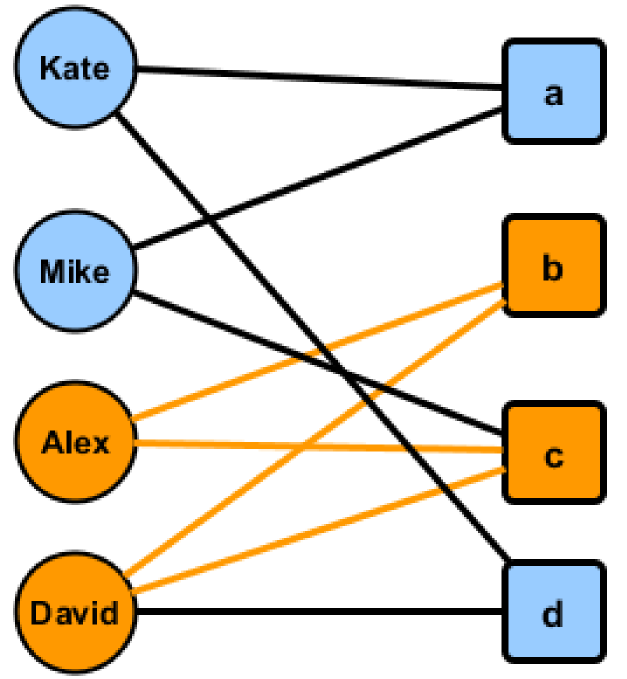
Example

	Car	**House**	**Laptop**	**Bicycle**
Kate	x			x
Mike	x		x	
Alex		x	x	
David		x	x	x

- *A pair ({Kate, Mike}*,{Car}) is a formal concept
- *({Alex, David} ,{Laptop}) doesn't form a formal concept, because {Laptop}$'$≠{Alex, David}*
- *({Alex, David} {House, Laptop}) is a formal concept*

FCA and Graphs

	a	b	c	d
Kate	x			x
Mike	x		x	
Alex		x	x	
David		x	x	x



Formal Context	Bipartite graph
Formal Cocept (maximal rectangle)	Biclique

Formal Concept Analysis

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 $\underline{\mathfrak{B}}(G,M,I)$ called ***concept lattice*** (***Galois lattice***).

Example

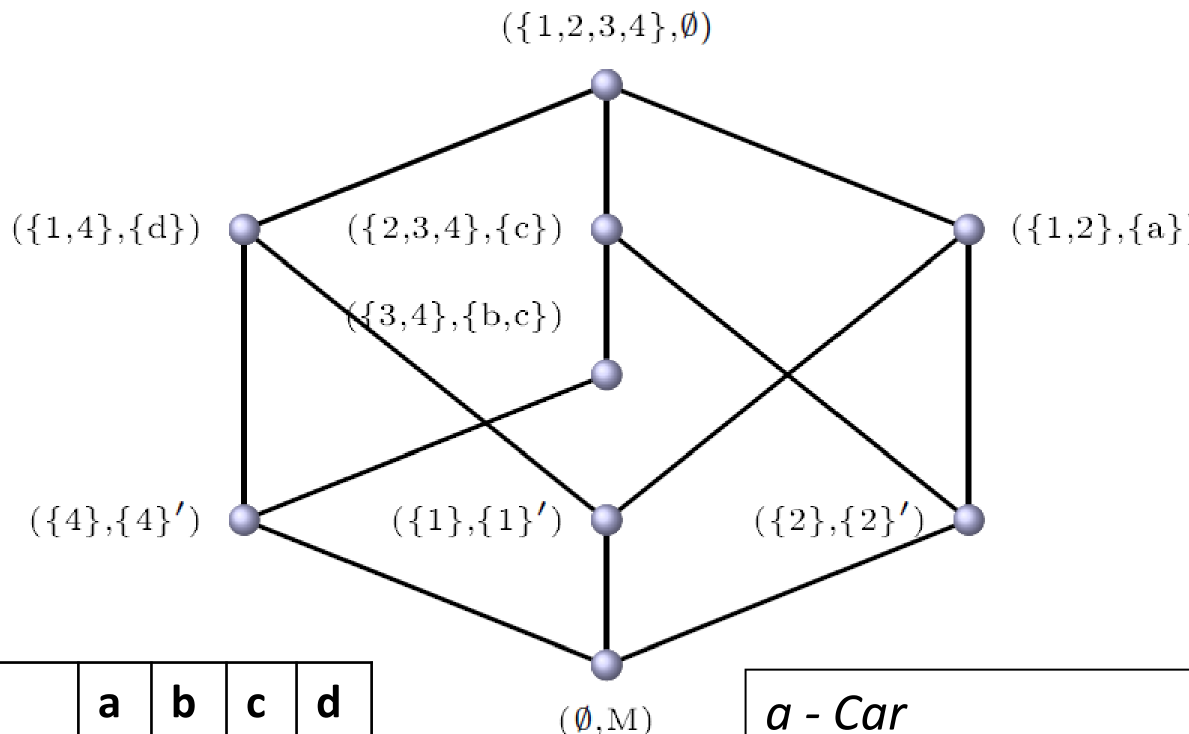
	Car	**House**	**Laptop**	**Bicycle**
Kate	x			x
Mike	x		x	
Alex		x	x	
David		x	x	x

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

is more general than concept

({Alex, David} {House, Laptop})

Concept Lattice Diagram



	a	b	c	d
Kate	x			x
Mike	x		x	
Alex		x	x	
David		x	x	x

a - Car

b - House

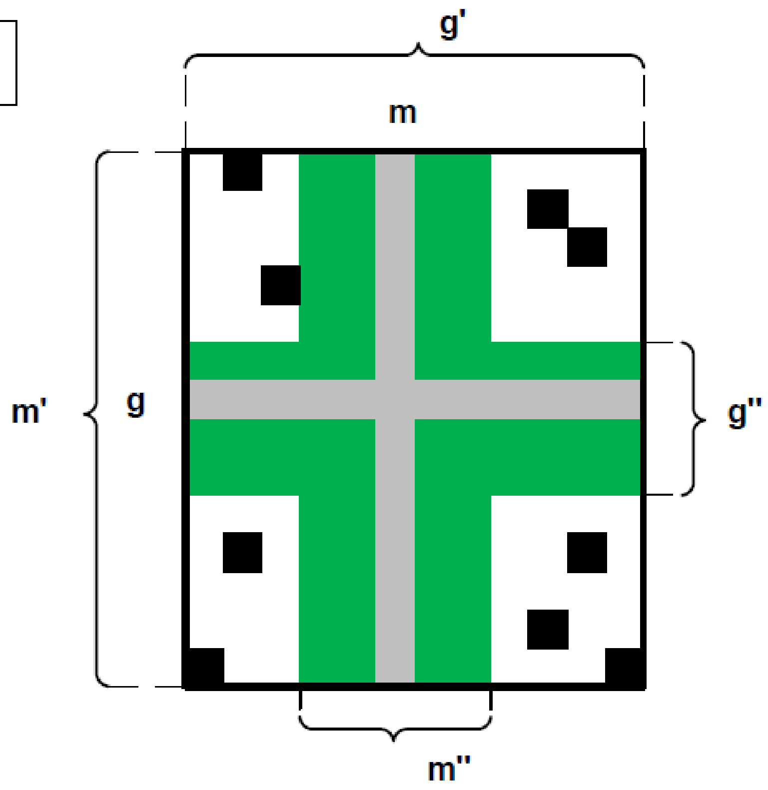
c - Laptop

d - Bicycle

Biclustering

Definition 1 If $(g, m) \in I$, then (m', g') is called an object-attribute or *oa-bicluster* with density $\rho(m', g') = \frac{|I \cap (m' \times g')|}{|m'| \cdot |g'|}$.

Geometrical inerpretation



Biclustering Example

	Car	**House**	**Laptop**	**Bicycle**
Kate	x			x
Mike	x		x	
Alex		x	x	
David		x	x	x

Since (House, David) is in the context

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

$\rho(House^{\prime}, David^{\prime})=5/6$

Biclustering properties

- Number of all biclusters for a context (G,M,I) not greater than $|I|$ vs $2^{\min\{|G|,|M|\}}$ 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.

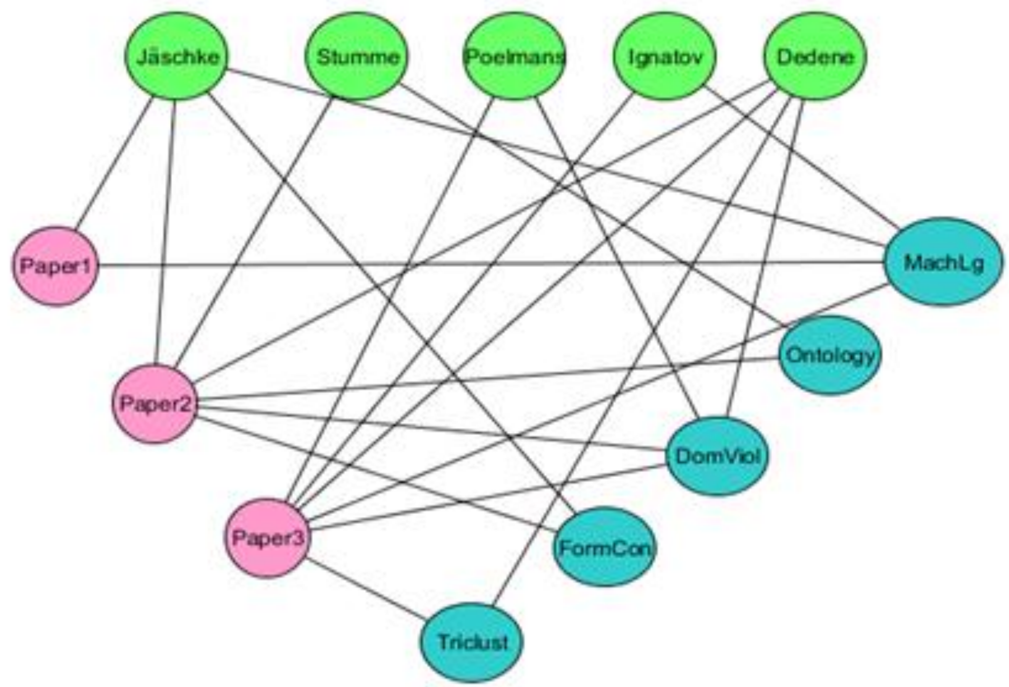
Triadic FCA and Folksonomies

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 $g \in G$ posseses 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

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

Prime operators of 1-sets	Their double prime counterparts
$m' = \{\ (g,b)\ \vert (g,m,b) \in Y\}$	$m'' = \{\ \tilde{m}\ \vert (g,b) \in m' \quad and \quad (g,\tilde{m},b) \in Y\}$
$g' = \{\ (m,b)\ \vert (g,m,b) \in Y\}$	$g'' = \{\ \tilde{g}\ \vert (m,b) \in g' \quad and \quad (\tilde{g},m,b) \in Y\}$
$b' = \{\ (g,m)\ \vert (g,m,b) \in Y\}$	$b'' = \{\ \tilde{b}\ \vert (g,m) \in b' \quad and \quad (g,m,\tilde{b}) \in Y\}$

To define triclusters we propose **box operators**

$$g^{\square} = \{\ g_i \mid (g_i, b_i) \in m' \ or \ (g_i, m_i) \in b'\}$$
$$m^{\square} = \{\ m_i \mid (m_i, b_i) \in g' \ or \ (g_i, m_i) \in b'\}$$
$$b^{\square} = \{\ b_i \mid (g_i, b_i) \in m' \ or \ (m_i, b_i) \in g'\}.$$

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 $T = (g^{\Box}, m^{\Box}, b^{\Box})$ 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 $\{t_1, t_2, t_3\}$

r_1

	t_1	t_2	t_3
u_1		×	×
u_2	×	×	×
u_3	×	×	×

r_2

	t_1	t_2	t_3
u_1	×	×	×
u_2	×		×
u_3	×	×	×

r_3

	t_1	t_2	t_3
u_1	×	×	×
u_2	×	×	×
u_3	×	×	

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

One dense tricluster VS 3^3 =27 formal triconcepts

Pseudo Triclustering for Social Networks

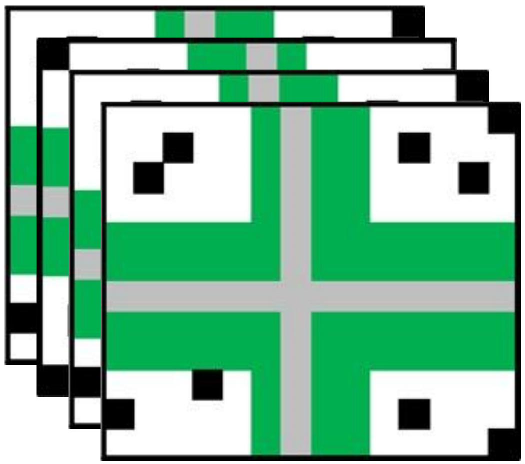
Let $K_{UI} = (U, I, X \subseteq U \times I)$ be a formal context which describes what interest $i \in I$ a particular user $u \in U$ has. Similarly, let $K_{UG} = (U, G, Y \subseteq U \times G)$ be a formal context which indicates what group $g \in G$ user $u \in U$ belongs to.

We can find dense biclusters as $(users, interesets)$ pairs in K_{UI} using oa-biclustering 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 $(users, groups)$.

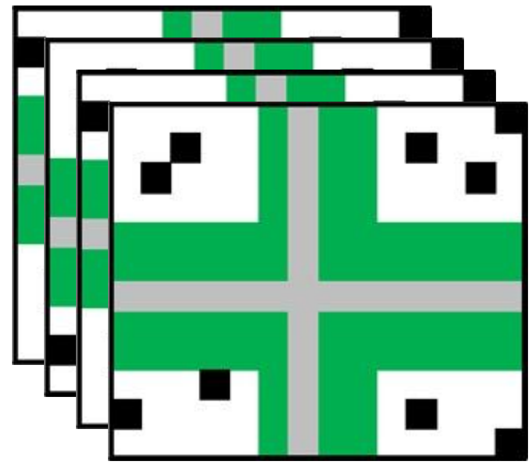
To this end we need to mine a (formal) tricontext $K_{UIG} = (U, I, G, Z \subseteq U \times I \times G)$, where (u, i, g) is in Z iff $(u, i) \in X$ and $(u, g) \in Y$. A particular tricluster has a form $T_k = (i^X \cap g^Y, u^X, u^Y)$ for every $(u, g, i) \in Z$ with $\frac{|i^X \cap g^Y|}{|i^X \cup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



$Set\ of\ biclusters$
$Bicl_{UI}$

$Set\ of\ biclusters$
$Bicl_{UG}$

Biclusters' extent similarity $\mu \geq C$

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

Algorithm

Let $Bicl_{UI}$ be a set of user-interest biclusters and $Bicl_{UG}$ be a set of user-group biclusters.

For each $(U_{ui}, I) \in Bicl_{UI}$ and $(U_{ug}, G) \in Bicl_{UG}$ 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, μ 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: $\rho \geq 0.5$.

Data

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:

	Bauman	MIPT	RSUH	RSSU
# users	18542	4786	10266	12281
# interests	8118	2593	5892	3733
# groups	153985	46312	95619	102046

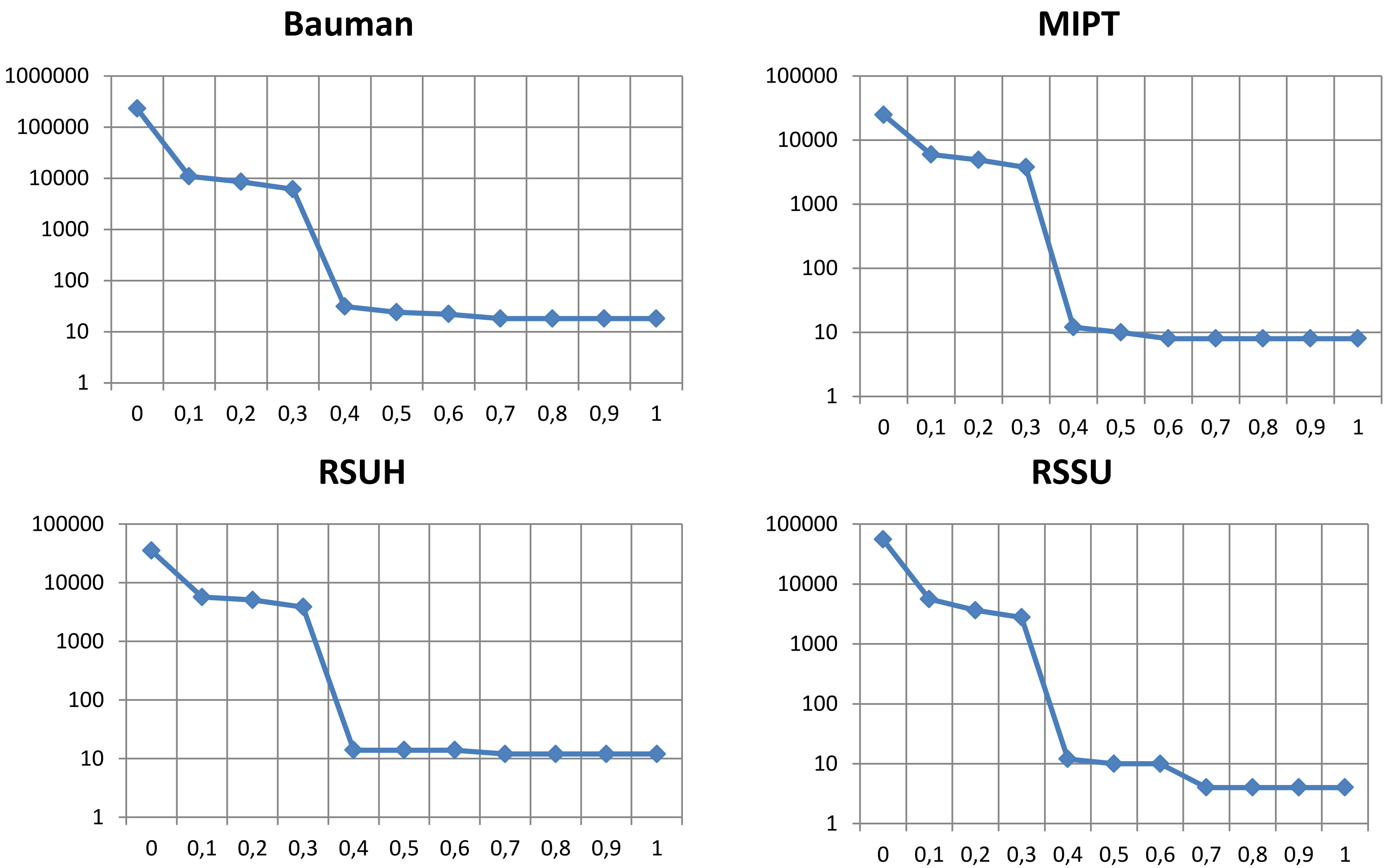
Biclustering results

ρ	Bauman				MIPT				RSUH				RSSU			
	UI		UG		UI		UG		UI		UG		UI		UG	
	Time	#	Time	#	Time	#	Time	#	Time	#	Time	#	Time	#	Time	#
0.0	9188	8863	1874458	248077	863	2492	109012	46873	3958	5293	519772	116882	2588	4014	693658	145086
0.1	8882	8331	1296056	173786	827	2401	91187	38226	3763	4925	419145	93219	2450	3785	527135	110964
0.2	8497	6960	966000	120075	780	2015	74498	28391	3656	4003	330371	68709	2369	3220	402159	79802
0.3	8006	5513	788008	85227	761	1600	63888	21152	3361	3123	275394	50650	2284	2612	332523	58321
0.4	7700	4308	676733	59179	705	1270	56365	15306	3252	2399	232154	35434	2184	2037	281164	40657
0.5	7536	3777	654047	53877	668	1091	54868	13828	3189	2087	224808	32578	2179	1782	270605	37244
0.6	7324	2718	522110	18586	670	775	44850	5279	3075	1367	174657	10877	2159	1264	211897	12908
0.7	7250	2409	511711	15577	743	676	43854	4399	3007	1224	171554	9171	2084	1109	208632	10957
0.8	7217	2326	508368	14855	663	654	43526	4215	3032	1188	170984	8742	2121	1081	209084	10503
0.9	7246	2314	507983	14691	669	647	43216	4157	2985	1180	174781	8649	2096	1072	206902	10422
1.0	7236	2309	511466	14654	669	647	43434	4148	3057	1177	173240	8635	2086	1068	207198	10408

Pseudo triclustering results

μ	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 μ

Examples. Biclusters

- ρ=83,33% Gen. pair: {3609, home}

G: {3609, 4566} M: {family, work, home}

- ρ =83,33% Gen. pair: {30568, orthodox church}

G: {25092, 30568} M: {music, monastery, orthodox church}

- ρ=100% Gen. pair: {4220, beauty}

G: {1269, 4220, 5337, 20787} M: {love, beauty}

Examples. Tricluster

- Measures:

μ : 100%;

Average ρ : 54,92%

Users: {16313, 24835}

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

Groups: {365, 457, 624,…, 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 intersts 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 potentionaly 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?