**Photo:** V. Batagelj, ***Araneus diadematus***

# Problems and projects

Vladimir Batagelj

University of Ljubljana

**Networks Workshop**

NICTA, Sydney, June 2005

version: June 28, 2005 / 12 : 36

# Outline

1 Open problems in GBM . . . . . 1
13 Further Readings / Books . . . . . 13
15 Further Readings / Papers . . . . . 15
17 Some links . . . . . 17



# Open problems in GBM

- GBM of valued networks
- GBM of multirelational networks
- GBM of temporal networks
- GBM of large networks

## GBM of valued networks

Can the clustering with relational constraint and blockmodeling problem be generalized to a common problem?

Batagelj V., Ferligoj A.: Clustering relational data. Data Analysis (ed.: W. Gaul, O. Opitz, M. Schader), Springer, Berlin 2000, 3-15.

## General problem of clustering relational data

The relationally constrained clustering problem with simple criterion function considers only the diagonal blocks that should be of one of the types $\Phi^i(R)$ . It also takes into account the dissimilarity matrix on units (derived from attribute data).

The blockmodeling problem deals only with relational data. The proposed optimization approach essentially expresses the constraints with a penalty function.

Both problems can be expressed as special cases of a clustering problem with a general criterion function of the form

- **G1s**. $P(\mathbf{C}) = \sum_{(C_1,C_2)\in\mathbf{C}\times\mathbf{C}} q(C_1, C_2)$, or
- **G1m**. $P(\mathbf{C}) = \max_{(C_1,C_2)\in\mathbf{C}\times\mathbf{C}} q(C_1, C_2)$

where $q$ is a block error satisfying

- **G2**. $q(C_1, C_2) \geq 0$

## . . . clustering relational data

The set of feasible clusterings $\Phi_k(R)$ for this problem is determined by the relation $R$ and additional requirements, such as:

- the blocks should be of selected types
- the model graph should be of specified form (prespecified)
- selected units should / should not be in the same cluster
- selected unit should / should not be in the selected cluster

## Approaches to the problem

There are different types of relational data (valued networks). In the following we shall assume $\mathcal{N} = (V, R, a, b)$ where $a : V \rightarrow A$ assigns a value to each unit/vertex and $b : R \rightarrow B$ assigns a value to each arc (link) of $R$. $A$ and $B$ are sets of values.

The function $b$ determines a matrix $\mathbf{B} = [b_{ij}]_{n \times n}, b_{ij} \in B \cup \{0\}$ and $b_{ij} = 0$ if units $i$ and $j$ are not connected by an arc.

We can approach the problem of clustering relational data by *indirect* (transformation to standard data analysis problems) or *direct* (formulating the problem as an optimization problem and solving it) approach.

## Indirect approach

A scenario for the indirect approach is to transform attribute data $a$ into dissimilarity matrix $D_a$ and network data $b$ into dissimilarity matrix $D_b$ and build criterion functions $P_a$ and $P_b$ based on them (they can be defined also directly from $a$ and $b$). Then we apply the multicriteria relationally constrained clustering methods on these functions.

We can also first combine $D_a$ and $D_b$ into a joint matrix $D_{ab}$ and apply relationally constrained clustering methods on it.

In a special case, when $D_b$ is defined as some corrected dissimilarity (see Batagelj, Ferligoj, Doreian, 1992) between descriptions $\mathbf{b}(x) = [\mathbf{B}(x), \mathbf{B}^T(x)]$, the relational data are built into $D_b$ and we can apply on the combined matrix $D_{ab}$ all standard methods for analysis of dissimilarity matrices.

## Direct approach

Again there are different possibilities:

1. Structural approach: used in program MODEL (Batagelj, 1996): Important is the structure (relation). Determine the best clustering $\mathbf{C}$ and the corresponding model. On the basis of $a$, $b$ and the obtained model compute values of model connections.

2. Multicriteria approach: construct two criterion functions: one based on values, the second based on structure. Solve the obtained multicriteria problem (Ferligoj, Batagelj, 1992).

3. Implicit approach: the types of connections are built into the criterion function combined with values.

Only the last approach needs some further explanations.

## Implicit approach

Let $\approx$ be an equivalence over set of units $V$, and $\mathcal{T}$ given types. We construct on blocks deviation functions $\delta(C_1, C_2; T), T \in \mathcal{T}$ such that $\approx$ is compatible with $\mathcal{T}$ over the network $\mathcal{N}$ iff

$$\forall u, v \in V \exists \delta(.,.;T), T \in \mathcal{T} : \delta([u],[v];T) = 0$$

Applying also an adequate normalization of $\delta$s we can construct a criterion function

$$P(\mathbf{C}) = \sum_{C_1, C_2 \in \mathbf{C}} \min_{T \in \mathcal{T}} \delta(C_1, C_2; T)$$

Evidently, $P(\mathbf{C}) = 0 \Leftrightarrow \mathbf{C}$ is compatible with $\mathcal{T}$ — all blocks of $\mathbf{C}$ are compatible with $\mathcal{T}$.

## Some examples

Assume that $a$ and $b$ are transformed into a matrix $\mathbf{A} = [a_{uv}]_{V \times V}$, $a_{uv} \geq 0$. Then

$$\begin{aligned}
\delta(X, Y; \mathrm{nul}) &= \frac{\sum_{x \in X, y \in Y} a_{xy}}{|X| \cdot |Y| \cdot \max\{a_{xy} : x \in X, y \in Y\}} \\
\delta(X, Y; \mathrm{rdo}) &= 1 - \max_{x \in X} \frac{\sum_{y \in Y} a_{xy}}{|Y| \cdot \max\{a_{xy} : y \in Y\}} \\
\delta(X, Y; \mathrm{cre}) &= 1 - \frac{\sum_{y \in Y} \max_{x \in X} a_{xy}}{|Y| \cdot \max\{a_{xy} : x \in X, y \in Y\}}
\end{aligned}$$

If $\max$ in the denominator equals to $0$ then also the fraction has value $0$.

Aleš Žiberna is working on approaches to GBM of valued networks.

## GBM of multirelational networks

Multiple networks are networks with more than one relation defined on the same set of vertices.

Assume that we have relations $R_1, R_2, \ldots, R_s$ and corresponding criterion functions $P_1, P_2, \ldots, P_s$ compatible respectively with $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_s$. Then we get the following multicriteria optimization problem: determine $\mathbf{C}^\star \in \Phi$ that 'minimizes'

$$(\Phi, P_1, P_2, \ldots, P_s)$$

This problem can be approched by (Ferligoj, Batagelj, 1992)

- multicriteria optimization. The solutions are Pareto points.
- transformation to single criterion optimization
  - by combining criterion functions: $\sum \alpha_i P_i$ or $\max \alpha_i P_i$, where $\alpha_i \geq 0$ and $\sum \alpha_i = 1$;
  - by combining relations.

## GBM of temporal networks

If the clustering $\mathbf{C}$ is the same for all time points we can treat the GBM of temporal networks problem as a GBM of multirelational networks for relations determined by the time slices.

In general, however, also the clustering $\mathbf{C}$ can change through time – the solution is a sequence of clusterings $(\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_s)$, $\mathbf{C}_i \in \Phi_i$ (or $\mathbf{C}_i \in \Phi_i(\mathbf{C}_{i-1})$).

The affect data from Sampson (1968) provide a useful source for examining structural balance through time (see Doreian and Mrvar, 1996).

## GBM of large networks

Large networks are usually sparse $m << n^2$.

In the case of given partition $\mathbf{C}$ (for example partition of contries to continents, patents to categories, ...) it is easy to determine the corresponding GBM that minimizes the criterion function.

A special class of GBM problems are symmetric-acyclic decompositions (Doreian, Batagelj, Ferligoj, 2000) for which also an algorithm for large networks was developed.

## . . . GBM of large networks

For some GBM problems on large networks the clustering methods can be used:

- if we have to compute the dissimilarities between attribute data for vertices we may consider methods that require only dissimilarities between the linked vertices;
- if we base the clustering on the descriptions of vertices using selected structural properties we can consider some variant of leader's method.

The leader's method works in $\Phi \times \Psi$.

$R = R_0 \in \Psi$
**repeat**
$C = H(R); R = G(C)$
**until** stabilizes;

Idea: $\Psi$ is the set feasible models.

# Further Readings / Books


- W. de Nooy, A. Mrvar, V. Batagelj: *Exploratory Social Network Analysis with Pajek*, CUP, 2005. Amazon. ESNA page.
- P. Doreian, V. Batagelj, A. Ferligoj: *Generalized Blockmodeling*, CUP, 2004. Amazon.
- P. Doreian, V. Batagelj, A. Ferligoj: *Positional analyses of sociometric data*. in Carrington, P.J., Scott, J., Wasserman, S., (Eds.) Models and Methods in Social Network Analysis. CUP, Berlin 2005, p. 77-97. Amazon.
- V. Batagelj, A. Mrvar: *Pajek – Analysis and Visualization of Large Networks*. in Jünger, M., Mutzel, P., (Eds.) Graph Drawing Software. Springer, Berlin 2003, p. 77-103. Amazon.
- S. Wasserman, K. Faust: *Social Network Analysis: Methods and Applications*. CUP, 1994. Amazon.
- J. P Scott: *Social Network Analysis: A Handbook*. SAGE Publications, 2000. Amazon.
- A. Degenne, M. Forsé: *Introducing Social Networks*. SAGE Publications, 1999. Amazon.
- U. Brandes, T. Erlebach (Eds.): *Network Analysis : Methodological Foundations*. LNCS, Springer, Berlin 2005. Amazon.

- S.N. Dorogovtsev, J.F.F. Mendes: *Evolution of Networks: From Biological Nets to the Internet and Www*. OUP, 2003. Amazon.
- P. Baldi, P. Frasconi, P. Smyth: *Modeling the Internet and the Web: Probabilistic Methods and Algorithms*. John Wiley & Sons, 2003. Amazon.
- J. Abello, P.M. Pardalos, M.G. Resende(Eds.): *Handbook of Massive Data Sets*. Springer, 2002. Amazon.
- R. Sedgewick, M. Schidlowsky: *Graph Algorithms (Algorithms in Java, Part 5)*, Third Edition. Addison-Wesley, 2003. Amazon.
- R.K. Ahuja, T.L. Magnanti, J.B. Orlin *Network Flows: Theory, Algorithms, and Applications*. Prentice Hall, 1993. Amazon.
- D.E. Knuth: *The Stanford GraphBase: A Platform for Combinatorial Computing*. Addison-Wesley, 1993. Amazon.
- F. Harary, R.Z. Norman, D. Cartwright: *Structural Models: An Intoroduction to the Theory of Directed Graphs*. John Wiley, 1965.
- F.S. Roberts: *Discrete Mathematical Models with Applications to Social, Biological, and Environmental Problems*. Prentice Hall, 1976. Amazon.

# Further Readings / Papers


- J. Kleinberg: *Authoritative sources in a hyperlinked environment*. Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1998. PDF.
- U. Brandes: *A Faster Algorithm for Betweenness Centrality*. Journal of Mathematical Sociology 25(2):163-177, 2001. PDF.
- V. Batagelj, U. Brandes: *Efficient Generation of Large Random Networks*. Physical Review E 71, 036113, 2005. PDF.
- H. Stuckenschmidt, M. Klein: *Structure-Based Partitioning of Large Concept Hierarchies*. Proceedings of the 3rd International Semantic Web Conference ISWC 2004, Hiroshima, Japan. PDF.
- J.I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, A. Vespignani: *k-core decomposition: a tool for the visualization of large scale networks*. cs.NI/0504107, 28 Apr 2005. PDF.
- W.W. Powell, D.R. White, K.W. Koput, J. Owen-Smith: *Network Dynamics and Field Evolution: The Growth of Interorganizational Collaboration in the Life Sciences*. American Journal of Sociology 110(4):1132-1205, 2005. PDF.
- V. Batagelj, A. Mrvar: *Density based approaches to network analysis: Analysis of Reuters terror news network*. LinkKDD2003, Washington, August 27, 2003. PDF.

- P. Doreian, V. Batagelj, A. Ferligoj: *Symmetric-Acyclic Decompositions of Networks*. Journal of Classification, 17(1), 3-28, 2000. PDF.
- V. Batagelj, A. Mrvar, M. Zaveršnik: *Network Analysis of Texts*. Erjavec, T., Gros, J. (Eds.) Language Technologies, Ljubljana, p. 143-148, 2002. PDF.
- K. Hamberger, M. Houseman, E. Daillant, D.R. White, L. Barry: *Matrimonial ring structures*. Forthcoming in Mathematiques, informatique, et sciences humaines, 2005. PDF.
- A. Mrvar, V. Batagelj: *Relinking Marriages in Genealogies*. Metodološki zvezki, 1(2), 407-418, 2004. PDF.
- M.E.J. Newman: *The structure and function of complex networks*. SIAM Review 45, 167-256, 2003. PDF.
- K.K. Mane, K. Börner: *Mapping topics and topic bursts in PNAS*. PNAS 101: 5287-5290, 2004. PDF, PNAS.

# Some links

- International Network for Social Network Analysis – INSNA.
- R.A. Hanneman, M. Riddle: *Introduction to social network methods*. href.
- *Pajek Data Sets*. href.
- *The InfoVis CyberInfrastructure Software*. href.
- ***Advanced Course on Knowledge Discovery***, Ljubljana, Slovenia, June 27-July 5, 2005. href.
- *Last version of these slides*. href.