Network structure, metadata and the prediction of missing nodes

Tiago P. Peixoto

Universität Bremen, Germany & ISI Foundation, Italy

Darko Hric

Aalto University, Finland

Santo Fortunato

Indiana University, USA

Seoul, May 2016

Networks with metadata

Many network datasets contain *metadata*: Annotations that go beyond the mere adjacency between nodes.

Often assumed as indicators of topological structure, and used to *validate* community detection methods. A.k.a. “ground-truth”.

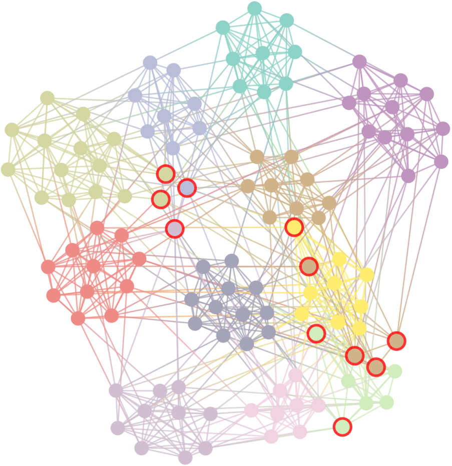
Example: American college football

Metadata (Conferences)

Example: American college football

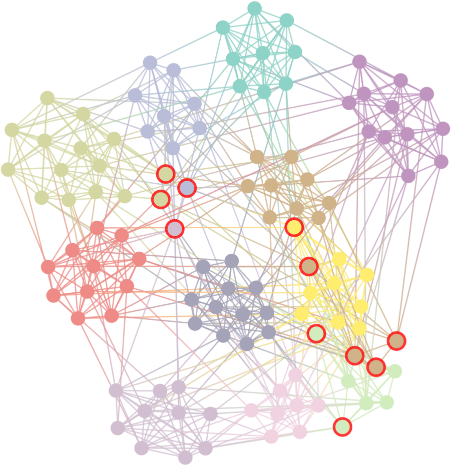
SBM fit

Example: American college football



Discrepancy

EXAMPLE: AMERICAN COLLEGE FOOTBALL

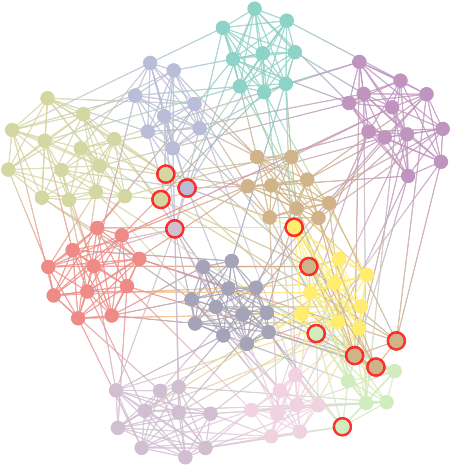


Discrepancy

Why the discrepancy?

Some hypotheses:

Example: American college football



Discrepancy

Why the discrepancy?

Some hypotheses:

- The model is not sufficiently descriptive.

Example: American college football

Discrepancy

Why the discrepancy?

Some hypotheses:

- The model is not sufficiently descriptive.
- The metadata is not sufficiently descriptive or is inaccurate.

EXAMPLE: AMERICAN COLLEGE FOOTBALL

Discrepancy

Why the discrepancy?

Some hypotheses:

- The model is not sufficiently descriptive.
- The metadata is not sufficiently descriptive or is inaccurate.
- Both.

EXAMPLE: AMERICAN COLLEGE FOOTBALL

Discrepancy

Why the discrepancy?

Some hypotheses:

- The model is not sufficiently descriptive.
- The metadata is not sufficiently descriptive or is inaccurate.
- Both.
- Neither.

Metadata is often very heterogeneous

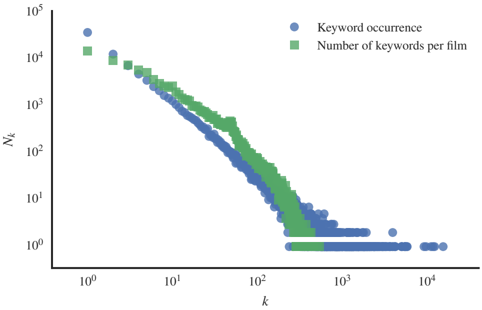
Example: IMDB Film-Actor network

Data: 96, 982 Films, 275, 805 Actors, 1, 812, 657 Film-Actor Edges

Film metadata: Title, year, genre, production company, country, user-contributed keywords, etc.

Actor metadata: Name, Age, Gender, Nationality, etc.

User-contributed keywords (93, 448)



Metadata is often very heterogeneous

Example: IMDB Film-Actor network

Keyword	Occurrences
'independent-film'	15513
'based-on-novel'	12303
'character-name-in-title'	11801
'murder'	11184
'sex'	9759
'female-nudity'	9239
'nudity'	5846
'death'	5791
'husband-wife-relationship'	5568
'love'	5560
'violence'	5480
'police'	5463
'father-son-relationship'	5063

Metadata is often very heterogeneous

Example: IMDB Film-Actor network

Keyword	Occurrences
’independent-film’	15513
’based-on-novel’	12303
’character-name-in-title’	11801
’murder’	11184
’sex’	9759
’female-nudity’	9239
’nudity’	5846
’death’	5791
’husband-wife-relationship’	5568
’love’	5560
’violence’	5480
’police’	5463
’father-son-relationship’	5063

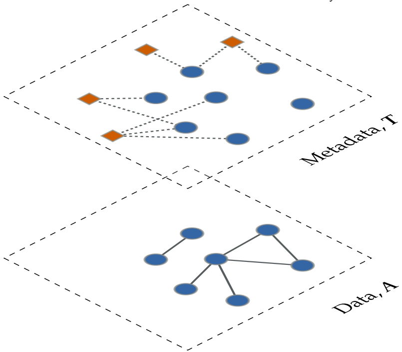
Keyword	Occurrences
’discriminaton-against-anteaters’	1
’partisan-violence’	1
’deliberately-leaving-something-behind’	1
’princess-from-outer-space’	1
’reference-to-aleksei-vorobyov’	1
’dead-body-on-the-beach’	1
’liver-failure’	1
’hit-with-a-skateboard’	1
’helping-blind-man-cross-street’	1
’abandoned-pet’	1
’retired-clown’	1
’resentment-toward-stepson’	1
’mutilating-a-plant’	1

BETTER APPROACH: METADATA AS DATA

Main idea: Treat metadata as data, not “ground truth”.

<u>Generalized annotations</u>

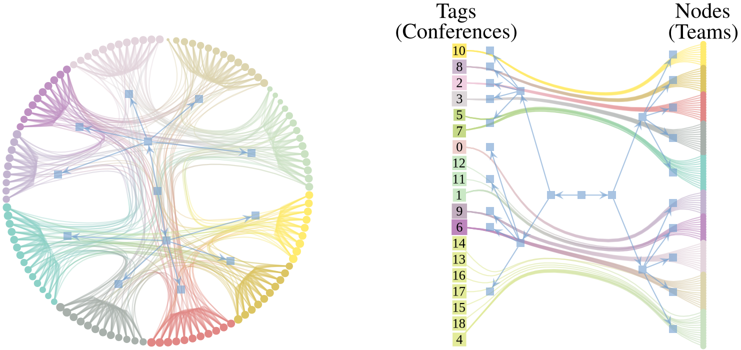
$A_{ij} \to$ Data layer
$T_{ij} \to$ Annotation layer



- Joint model for data and metadata (the layered SBM [1]).
- Arbitrary types of annotation.
- Both data and metadata are clustered into groups.
- Fully nonparametric.

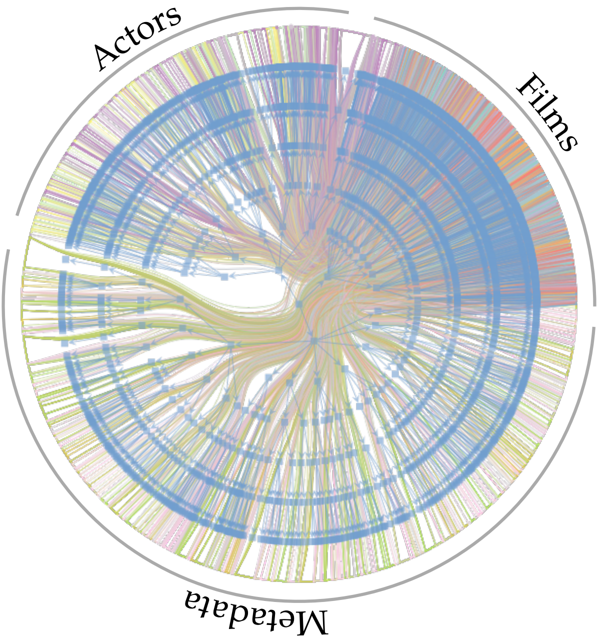
[1] T.P.P, Phys. Rev. E 92, 042807 (2015)

Example: American college football



(a) Data (b) Metadata

Example: IMDB film-actor network



PREDICTION OF MISSING EDGES

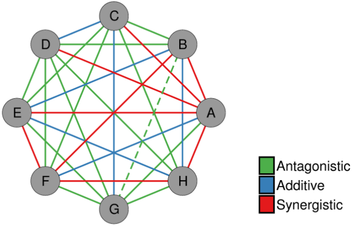
$$G' = \underbrace{G}_{\text{Observed}} \cup \underbrace{\delta G}_{\text{Missing}}$$

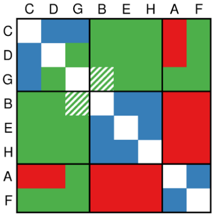
Posterior probability of missing edges

$$P(\delta G|G,\{b_i\}) = \frac{\sum_\theta P(G \cup \delta G|\{b_i\},\theta)P(\theta)}{\sum_\theta P(G|\{b_i\},\theta)P(\theta)}$$

A. Clauset, C. Moore, MEJ Newman, Nature, 2008
R. Guimerà, M Sales-Pardo, PNAS 2009

Drug-drug interactions





R. Guimerà, M. Sales-Pardo, PLoS Comput Biol, 2013

METADATA AND PREDICTION OF *missing nodes*

Node probability, with known group membership:

$$P(\boldsymbol{a}_i|\boldsymbol{A}, b_i, \boldsymbol{b}) = \frac{\sum_\theta P(\boldsymbol{A}, \boldsymbol{a}_i|b_i, \boldsymbol{b}, \theta)P(\theta)}{\sum_\theta P(\boldsymbol{A}|\boldsymbol{b}, \theta)P(\theta)}$$

Node probability, with unknown group membership:

$$P(\boldsymbol{a}_i|\boldsymbol{A}, \boldsymbol{b}) = \sum_{b_i} P(\boldsymbol{a}_i|\boldsymbol{A}, b_i, \boldsymbol{b})P(b_i|\boldsymbol{b}),$$

Node probability, with unknown group membership, but known metadata:

$$P(\boldsymbol{a}_i|\boldsymbol{A}, \boldsymbol{T}, \boldsymbol{b}, \boldsymbol{c}) = \sum_{b_i} P(\boldsymbol{a}_i|\boldsymbol{A}, b_i, \boldsymbol{b})P(b_i|\boldsymbol{T}, \boldsymbol{b}, \boldsymbol{c}),$$

Group membership probability, given metadata:

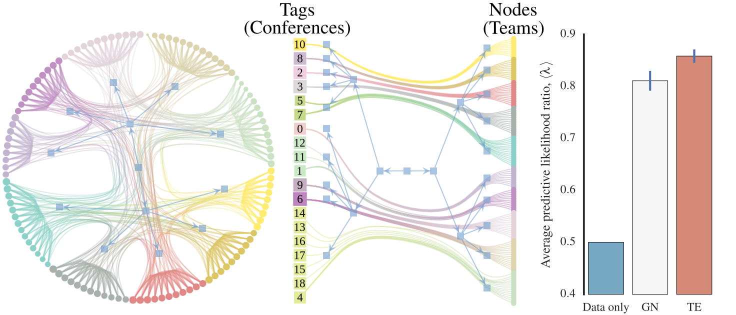
$$P(b_i|\boldsymbol{T}, \boldsymbol{b}, \boldsymbol{c}) = \frac{P(b_i, \boldsymbol{b}|\boldsymbol{T}, \boldsymbol{c})}{P(\boldsymbol{b}|\boldsymbol{T}, \boldsymbol{c})} = \frac{\sum_\gamma P(\boldsymbol{T}|b_i, \boldsymbol{b}, \boldsymbol{c}, \gamma)P(b_i, \boldsymbol{b})P(\gamma)}{\sum_{b_i'} \sum_\gamma P(\boldsymbol{T}|b_i', \boldsymbol{b}, \boldsymbol{c}, \gamma)P(b_i', \boldsymbol{b})P(\gamma)}$$

Predictive likelihood ratio:

$$\lambda_i = \frac{P(\boldsymbol{a}_i|\boldsymbol{A}, \boldsymbol{T}, \boldsymbol{b}, \boldsymbol{c})}{P(\boldsymbol{a}_i|\boldsymbol{A}, \boldsymbol{T}, \boldsymbol{b}, \boldsymbol{c}) + P(\boldsymbol{a}_i|\boldsymbol{A}, \boldsymbol{b})}$$

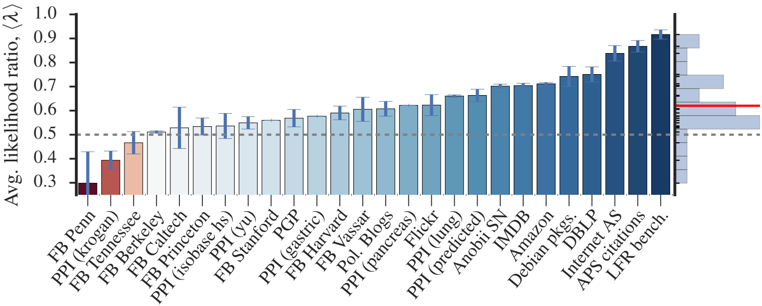
$\lambda_i > 1/2 \rightarrow$ the metadata improves the prediction task

Metadata and prediction of missing nodes



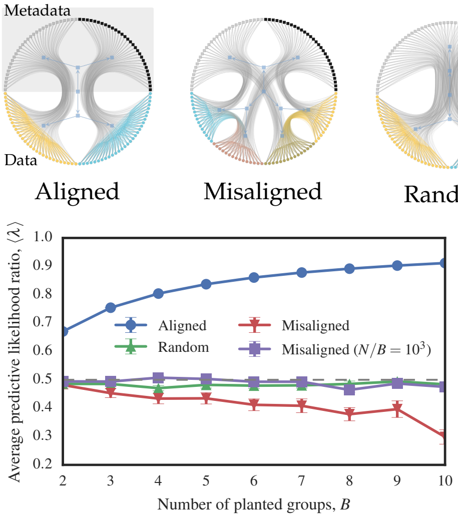
(a) Data (b) Metadata (c) Node prediction

Metadata and prediction of missing nodes



$$\lambda_i = \frac{P(\boldsymbol{a}_i|\boldsymbol{A},\boldsymbol{T},\boldsymbol{b},\boldsymbol{c})}{P(\boldsymbol{a}_i|\boldsymbol{A},\boldsymbol{T},\boldsymbol{b},\boldsymbol{c}) + P(\boldsymbol{a}_i|\boldsymbol{A},\boldsymbol{b})}$$

Metadata and prediction of missing nodes



Metadata predictiveness

Neighbor probability:

$$P_e(i|j) = k_i \frac{e_{b_i,b_j}}{e_{b_i} e_{b_j}}$$

Neighbour probability, given metadata tag:

$$P_t(i) = \sum_j P(i|j) P_m(j|t)$$

Null neighbor probability (no metadata tag):

$$Q(i) = \sum_j P(i|j) \Pi(j)$$

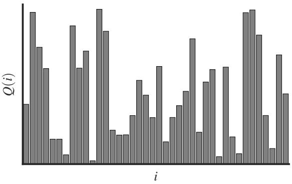
Kullback-Leibler divergence:

$$D_{\mathrm{KL}}(P_t||Q) = \sum_i P_t(i) \ln \frac{P_t(i)}{Q(i)}$$

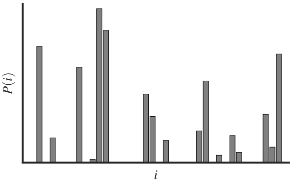
Relative divergence:

$$\mu_r \equiv \frac{D_{\mathrm{KL}}(P_t||Q)}{H(Q)} \rightarrow \text{Metadata group predictiveness}$$

Neighbour prob. without metadata

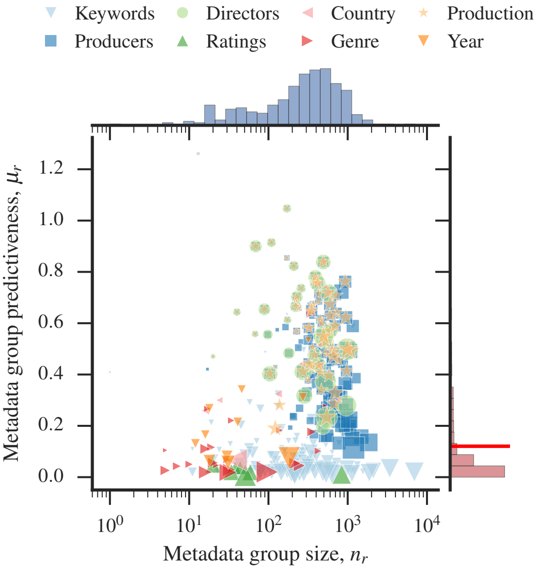


Neighbour prob. with metadata



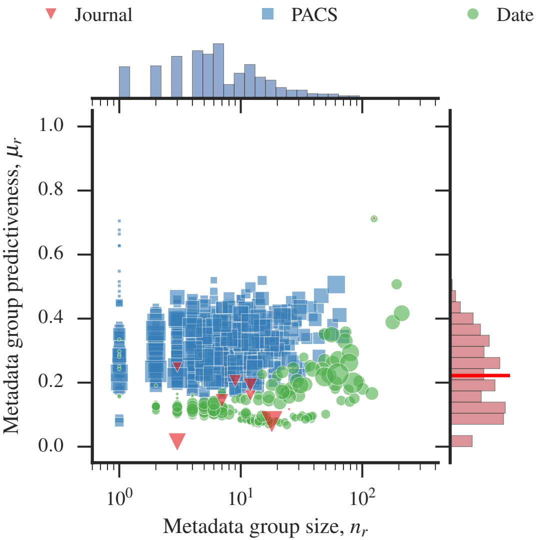
Metadata predictiveness

IMDB film-actor network



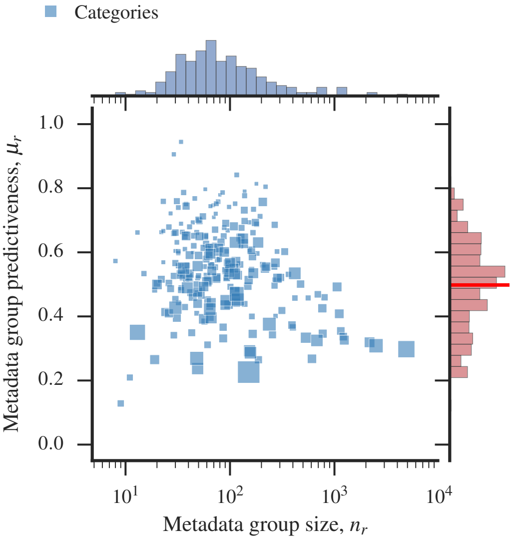
METADATA PREDICTIVENESS

APS CITATION NETWORK



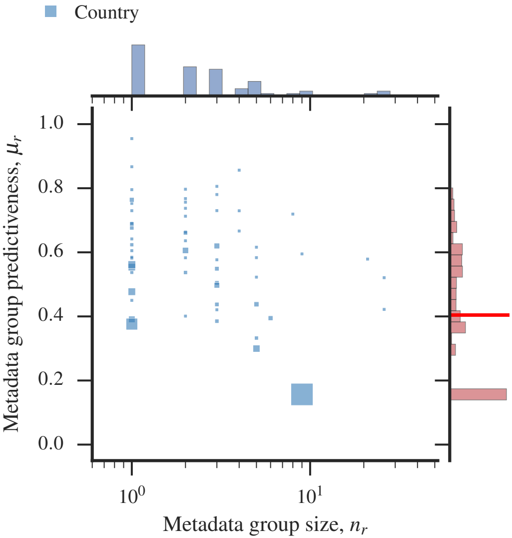
METADATA PREDICTIVENESS

AMAZON CO-PURCHASES



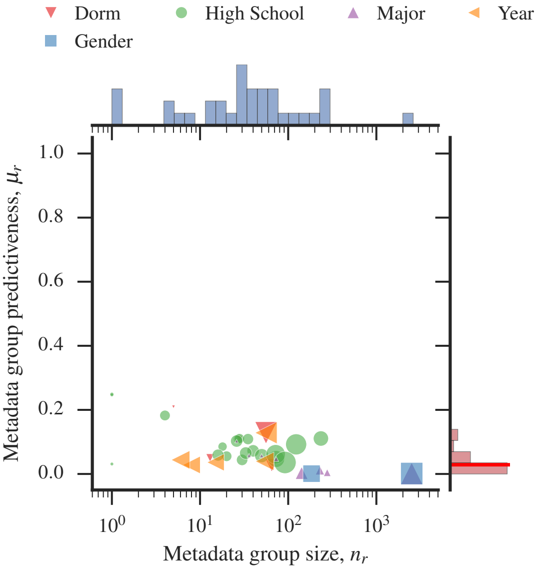
Metadata predictiveness

Internet AS



Metadata predictiveness

Facebook Penn State



THE END

Main Message:

- Metadata is often structured, heterogeneous and noisy.
- It is in general not trivially descriptive of network structure ($\neq$ "ground truth").
- It should be treated as part of the data, and modeled.

Darko Hric, T. P. P., Santo Fortunato, arXiv:1604.00255

Other talks:

"The Trouble with Community Detection"
M. E. J. Newman and Aaron Clauset
Wed. 14:00, Dongkang B, 3F

"The Ground Truth about Metadata and Community Detection in Networks"
Leto Peel, Daniel B. Larremore and Aaron Clauset
Wed. 15:00, Dongkang B, 3F

graph-tool

Very fast, freely available C++ code as part of the graph-tool Python library.
`http://graph-tool.skewed.de`

Efficient inference algorithms

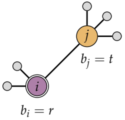
T. P. Peixoto, Phys. Rev. E 89, 012804 (2014)

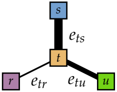
Smart MCMC

- Choose a random vertex v (happens to belong to block r).
- Move it to a random block $s \in [1, B]$, chosen with a probability $p(r \to s|t)$ proportional to $e_{ts} + \epsilon$, where t is the block membership of a randomly chosen neighbour of v.
- Accept the move with probability

$$a = \min\left\{ e^{-\beta\Delta S}\frac{\sum_t p_t^i p(s \to r|t)}{\sum_t p_t^i p(r \to s|t)}, 1 \right\}.$$

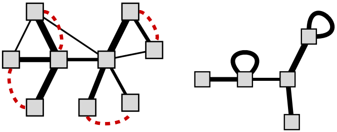
- Repeat.





Fast mixing times.

Agglomerative initialization



Avoids metastable states.

Algorithmic complexity:

$O(N \ln^2 N)$
(independent of B)

Scales up to $10^7 - 10^8$ edges.

graph-tool

Freely available efficient implementation
http://graph-tool.skewed.de