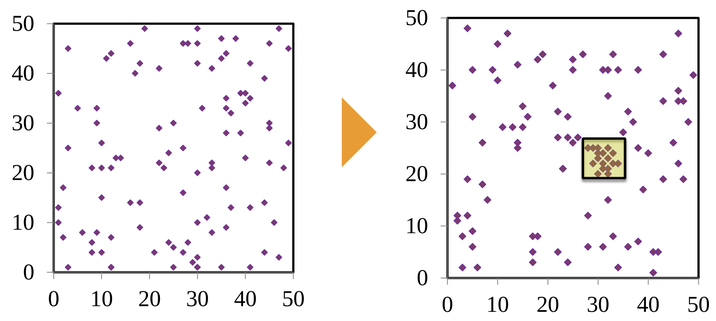
Beyond Blocks: Hyperbolic Community Detection

Miguel Araújo[1,2], Stephan Günnemann[1], Gonzalo Mateos[1] and Christos Faloutsos[1]
[1] *Carnegie Mellon University, Computer Science Department, USA*
[2] *Universidade do Porto, CRACS/INESC-TEC, Portugal*

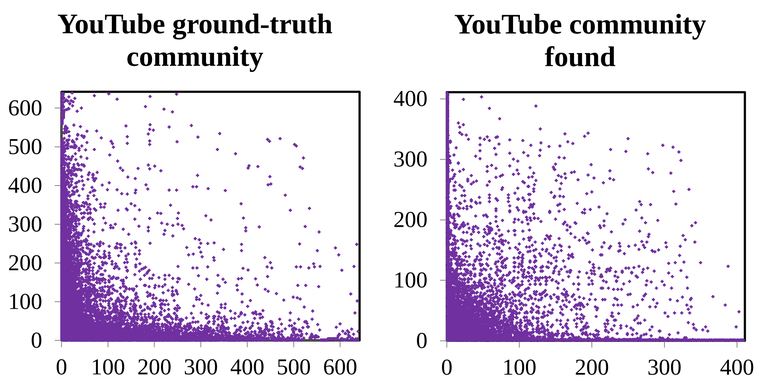
Why Community Detection?

- **Automatic Categorization –** articles on the same subject link to each other.
- **Recommendation Systems –** friends have similar interests.
- **Fraud Detection –** in bipartite reviews graph, dense subgraphs might indicate fraud.

Current Methods

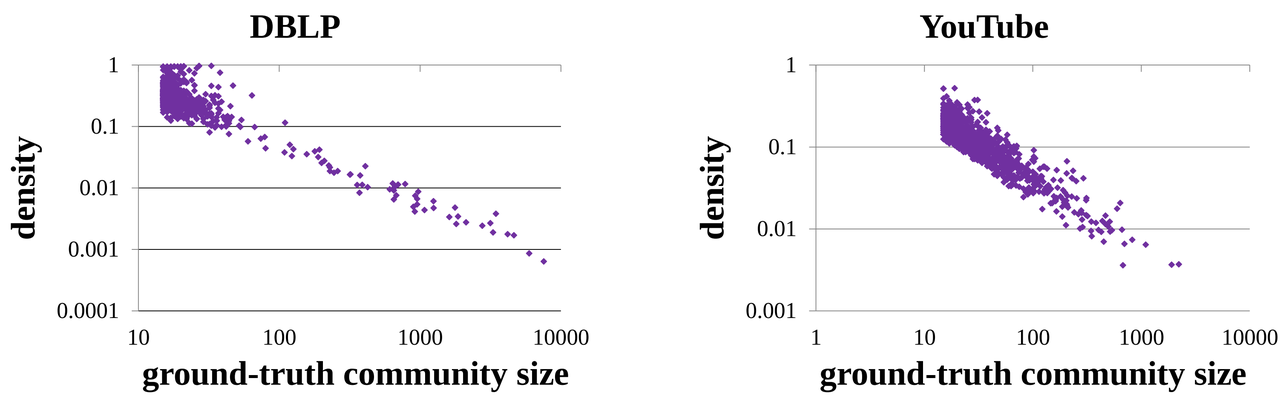


→ **We argue that communities have Hyperbolic shape**

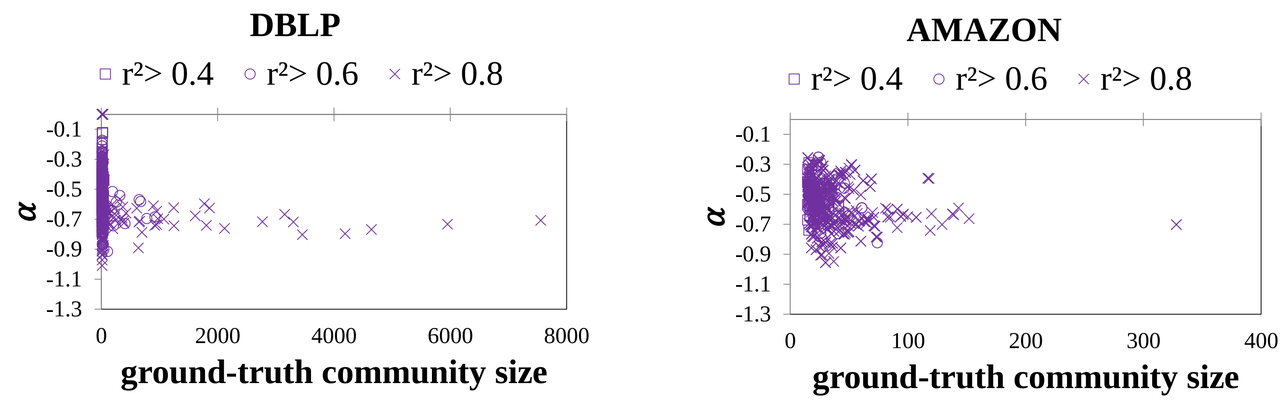


Empirical Evidence

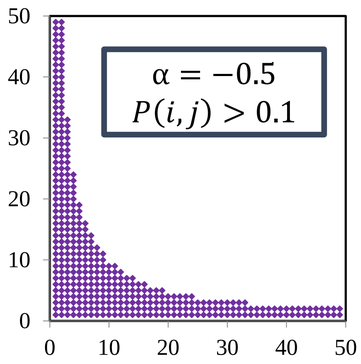
1. **Density decreases with size**



2. **Nodes are not homogeneous**
   Ground-truth communities seem to follow a power-law degree distribution: $degree_i \approx K \cdot rank_i^{\alpha}$



Hyperbolic Community Model



Assuming independence:

$$\boldsymbol{P(i,j) \propto d_i \cdot d_j}$$

$$\boldsymbol{P(i,j) \propto i^{\alpha} \cdot j^{\alpha}}$$

MDL principle

Find communities that minimize:
a. Cost of encoding the model
b. Cost of encoding the errors

Given a matrix $\mathbf{M} \in \{\mathbf{0}, \mathbf{1}\}^{N \times N}$, find $\boldsymbol{C}^* \subseteq (P(\mathcal{N}) \times \mathbb{R} \times \mathbb{R})$ with

$$\boldsymbol{C}^* = \arg\min_{\boldsymbol{C}} [\log^* |\boldsymbol{C}| + \textstyle\sum_{C \in \boldsymbol{C}} L_1(C) + L_2(\mathbf{M}|\mathbf{C})] \text{, where}$$

$$a)\ L_1(C_i) = \log N + |S_i| \cdot \log N + k_{\alpha_i} + \log |S_i|^2$$

$$b)\ L_2(\mathbf{M}|\mathbf{C}) = \log^* \left\|\mathbf{M} - \mathbf{M}_r^{\boldsymbol{C}}\right\|_F^2 + 2\left\|\mathbf{M} - \mathbf{M}_r^{\boldsymbol{C}}\right\|_F^2 \cdot \log N$$

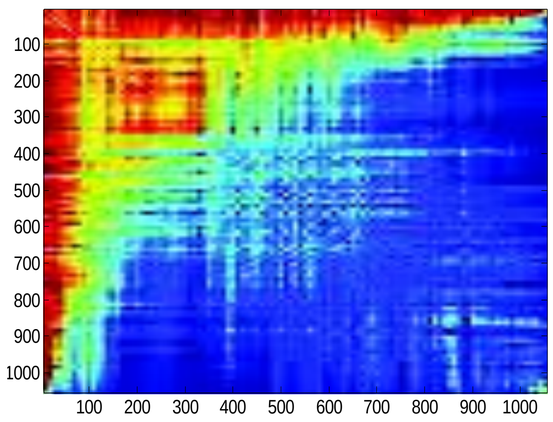
Efficient Algorithm

1. **Candidate community:** compute rank-1 decomposition to guide search;
2. **Minimize MDL objective** through sampling using rank-1 scores as bias.
3. **Deflate matrix** to find new communities.

→ **Scalable:** linear in the number of non-zeros and number of communities.

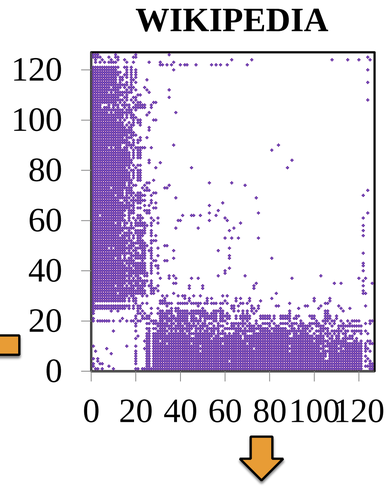
→ **No user-defined parameters:** fully automatic.

Experimental Results



LiveJournal community

- Country
- List_of_countries_by_area
- List_of_FIFA_country_codes
- Members_of_the_United_Nations
- List_of_national_rulers
- Gross_domestic_product
- …



Countries (e.g. Portugal, France, etc.)

Dataset	# of nodes	# of edges
AMAZON	334 863	1 851 744
DBLP	317 081	2 099 732
YOUTUBE	1 134 890	5 975 248
LIVEJOURNAL	3 997 962	34 681 889
WIKIPEDIA	143 508	3 753 156