How community-like is the structure of synthetically generated graphs?

Arnau Prat-Pérez
Universitat Politècnica de Catalunya
Barcelona

David Dominguez-Sal
Sparsity Technologies
Barcelona

DAMA-UPC. DATA MANAGEMENT
UNIVERSITAT POLITÈCNICA DE CATALUNYA

*Sparsity

LDBC

Motivation

- Community Detection is typically tested using **synthetic graphs (LFR generator)**.
  - Not only the graph output, but **communities** also.
- Recently, real graphs with ground truth have acquired popularity.
- How realistic is the community structure of synthetically generated graphs?
  - Existing work on vertex centric characteristics.

Methodology

- We select real datasets with **ground truth communities**.
- We select two synthetic generators: **LFR** and **LDBC Data Generator**.
  - They output communities.
- We select a set of **6 metrics**.
- For each pair of graphs and each metric, we compare the distributions of the communities using the **Spearman's correlation coefficient**.

Real Graphs

- Widely used in the literature.
- Diverse origin.
- Different sizes.

	Nodes	Edges
Amazon	334,863	925,872
Dblp	317,080	1,049,866
Youtube	1,134,890	2,987,624
LiveJournal	3,997,962	34,681,189

LFR Generator

- LFR
  - Generator created as a benchmark for Community Detection.
  - Five graphs with different mixing factors: 0.1 to 0.5.
  - Other parameters matching those found in real graphs.
  - Communities directly output by the program.

	Nodes	Edges
Lfr.1	150,000	649,538
Lfr.2	150,000	650,163
Lfr.3	150,000	650,946
Lfr.4	150,000	649,363
Lfr.5	150,000	648,128

LDBC Data Generator

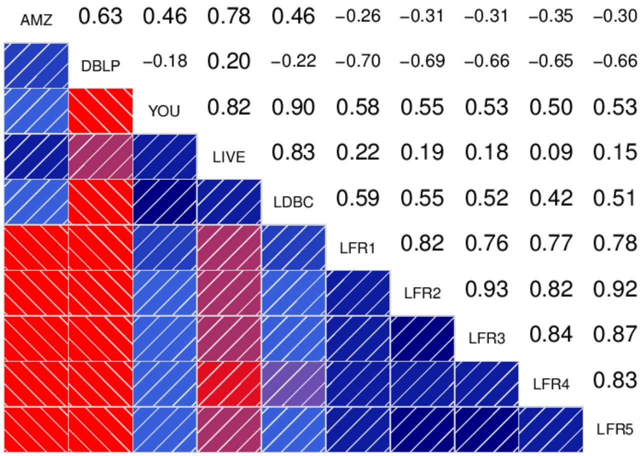
- LDBC Data Generator
  - Data Generator of the LDBC Social Network Benchmark.
  - Communities are created from metadata.
  - One instance, simulating 3 years of 150000 users activity.

	Nodes	Edges	Communities
LDBC	150,000	5,530,880	2,110,508

Metrics

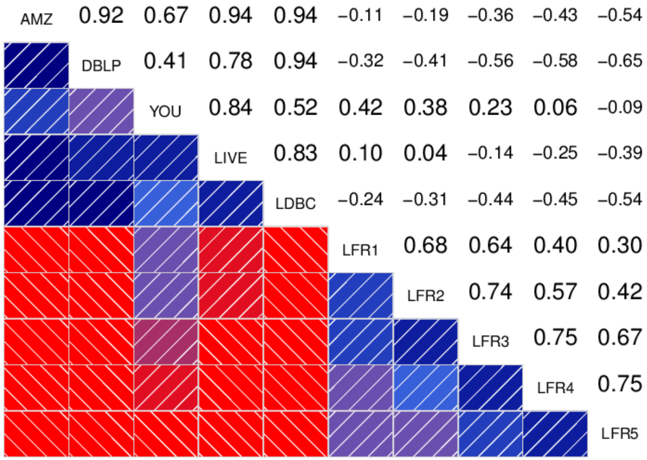
- 4 metrics for the internal structure:
  - Clustering Coefficient
  - Triangle Participation Ratio (TPR)
  - Bridge Ratio
  - Diameter
- 1 metric for the external connectivity.
  - Conductance
- Also the Size.

Correlations



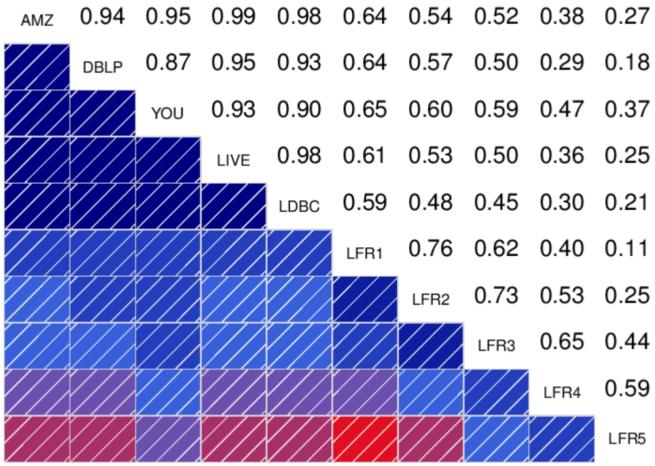
	DBLP	YOU	LIVE	LDBC	LFR1	LFR2	LFR3	LFR4	LFR5
AMZ	0.63	0.46	0.78	0.46	−0.26	−0.31	−0.31	−0.35	−0.30
DBLP		−0.18	0.20	−0.22	−0.70	−0.69	−0.66	−0.65	−0.66
YOU			0.82	0.90	0.58	0.55	0.53	0.50	0.53
LIVE				0.83	0.22	0.19	0.18	0.09	0.15
LDBC					0.59	0.55	0.52	0.42	0.51
LFR1						0.82	0.76	0.77	0.78
LFR2							0.93	0.82	0.92
LFR3								0.84	0.87
LFR4									0.83

Clustering Coefficient



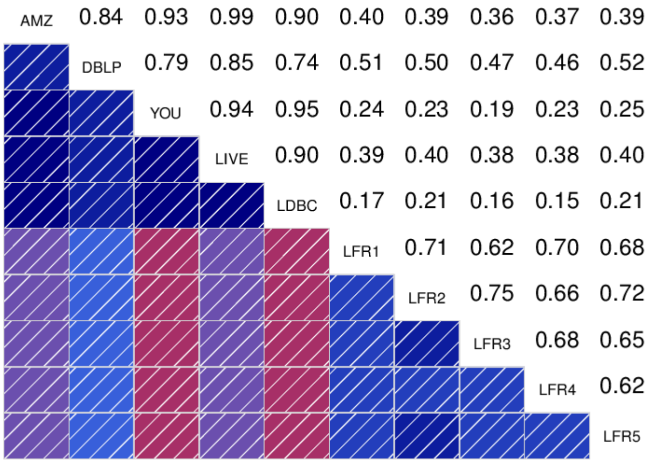
	DBLP	YOU	LIVE	LDBC	LFR1	LFR2	LFR3	LFR4	LFR5
AMZ	0.92	0.67	0.94	0.94	−0.11	−0.19	−0.36	−0.43	−0.54
DBLP		0.41	0.78	0.94	−0.32	−0.41	−0.56	−0.58	−0.65
YOU			0.84	0.52	0.42	0.38	0.23	0.06	−0.09
LIVE				0.83	0.10	0.04	−0.14	−0.25	−0.39
LDBC					−0.24	−0.31	−0.44	−0.45	−0.54
LFR1						0.68	0.64	0.40	0.30
LFR2							0.74	0.57	0.42
LFR3								0.75	0.67
LFR4									0.75

TPR



	DBLP	YOU	LIVE	LDBC	LFR1	LFR2	LFR3	LFR4	LFR5
AMZ	0.94	0.95	0.99	0.98	0.64	0.54	0.52	0.38	0.27
DBLP		0.87	0.95	0.93	0.64	0.57	0.50	0.29	0.18
YOU			0.93	0.90	0.65	0.60	0.59	0.47	0.37
LIVE				0.98	0.61	0.53	0.50	0.36	0.25
LDBC					0.59	0.48	0.45	0.30	0.21
LFR1						0.76	0.62	0.40	0.11
LFR2							0.73	0.53	0.25
LFR3								0.65	0.44
LFR4									0.59

Bridges Ratio



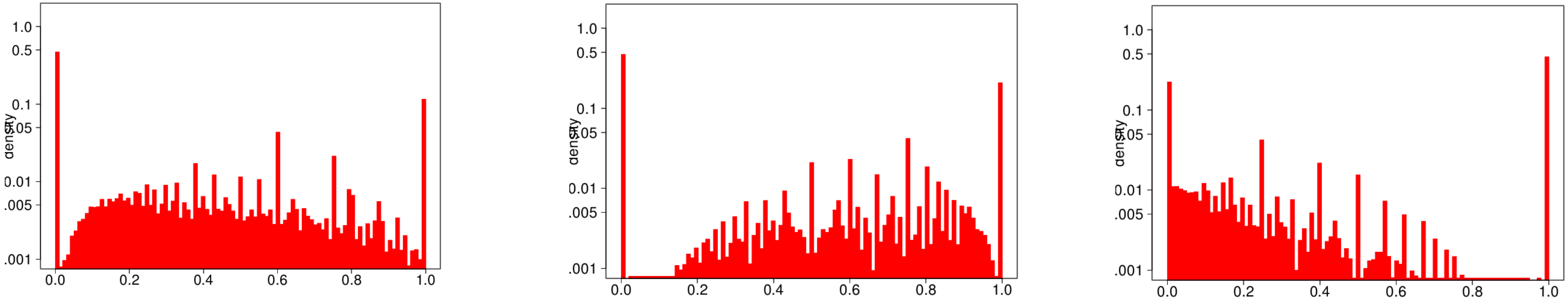
	DBLP	YOU	LIVE	LDBC	LFR1	LFR2	LFR3	LFR4	LFR5
AMZ	0.84	0.93	0.99	0.90	0.40	0.39	0.36	0.37	0.39
DBLP		0.79	0.85	0.74	0.51	0.50	0.47	0.46	0.52
YOU			0.94	0.95	0.24	0.23	0.19	0.23	0.25
LIVE				0.90	0.39	0.40	0.38	0.38	0.40
LDBC					0.17	0.21	0.16	0.15	0.21
LFR1						0.71	0.62	0.70	0.68
LFR2							0.75	0.66	0.72
LFR3								0.68	0.65
LFR4									0.62

Log10(Size)

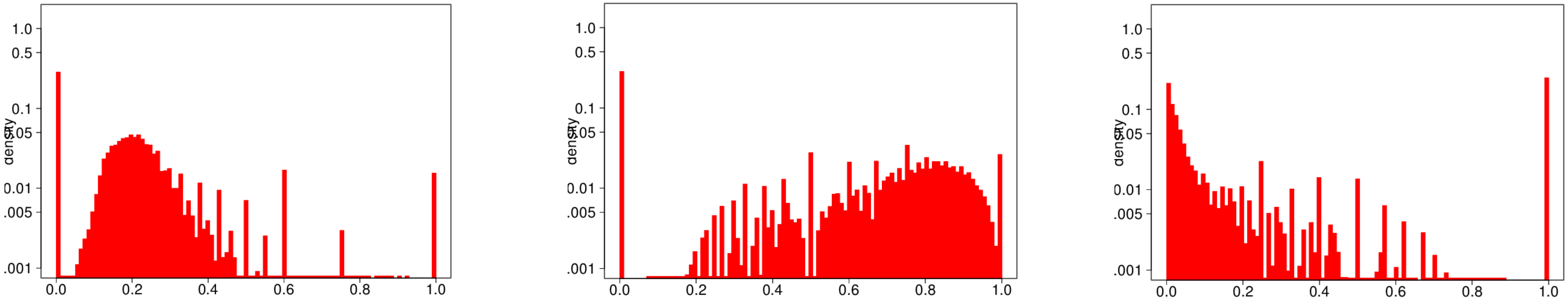
Multimodality

- Multimodal distributions for CC, TPR and Bridge Ratio.

LiveJournal



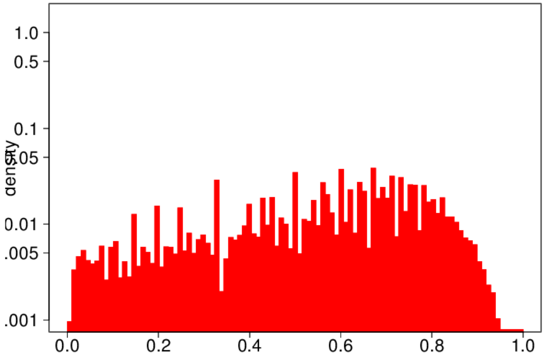
LDBC



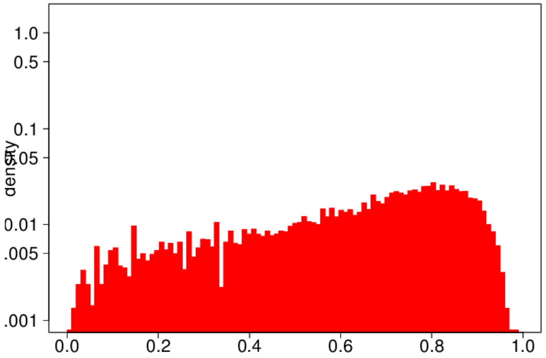
Clustering Coefficient TPR Bridge Ratio

Findings on real graphs

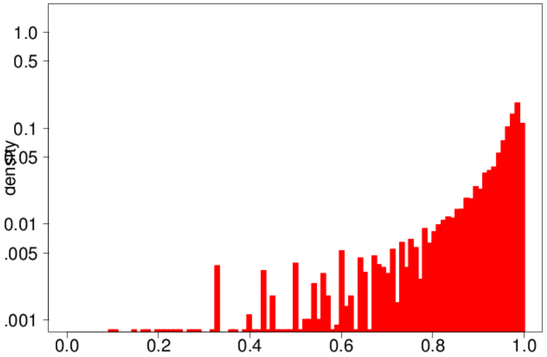
- Signs of two different **Conductance** profiles



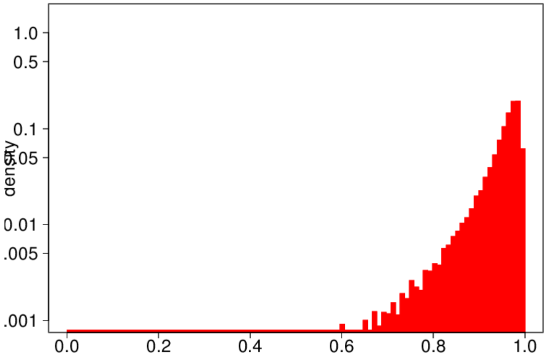
Amazon



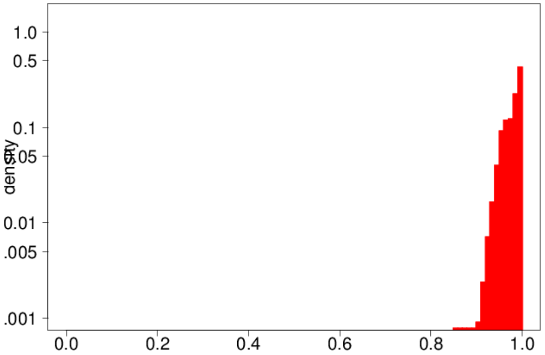
Dblp



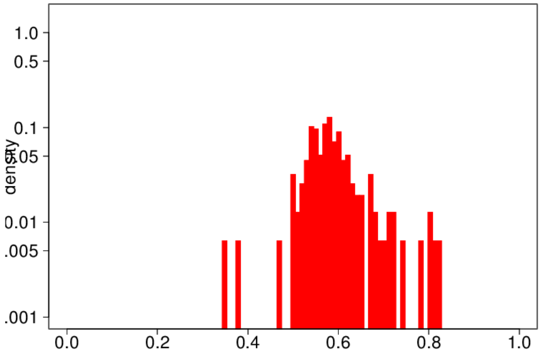
Youtube



Livejournal



LDBC



LFR3

Conclusions

- Real graphs show similar distributions.
- LDBC Data Generator distributions are more realistic than those produced by LFR.
- Some distributions are multimodal: LDBC Data Generator mimics this.
- Signs of two different conductance profiles.
- Future Work: Experiment with more parameter configurations.

Thank you!