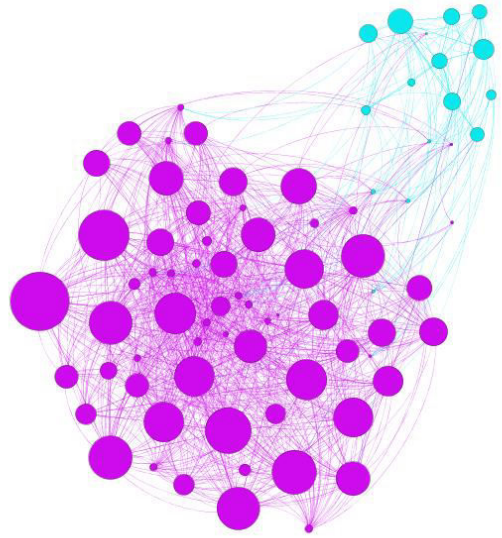
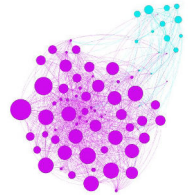
Project: VK network analysis

German Sokolov

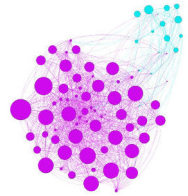


Outline



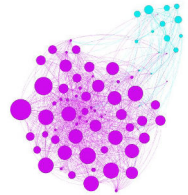
1. Network summary
2. Structural analysis
   - Centrality
   - Assortativity
   - Similarity
   - Approximating random graph
3. Community detection
   - Cliques
   - K-shells
   - Communities

Outline



1. Network summary
2. Structural analysis
   - Centrality
   - Assortativity
   - Similarity
   - Approximating random graph
3. Community detection
   - Cliques
   - K-shells
   - Communities

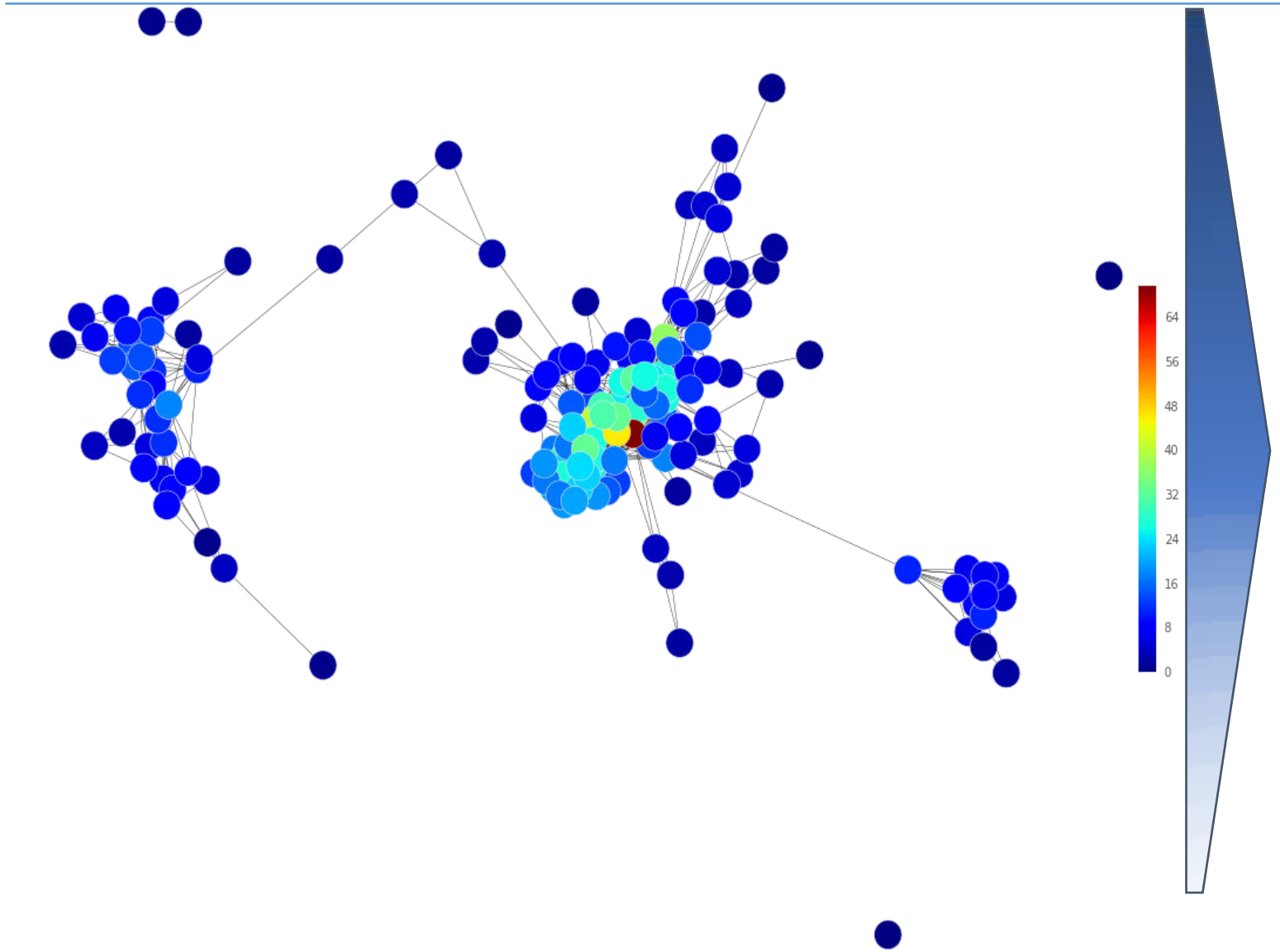
Network summary



- Data collection – VK API (“vk” library)
- Attributes of nodes*:
  - Name
  - Sex
  - City
  - University
- Key informaton:
  - Number of nodes - 162
  - Number of edges – 1 046
  - Diameter - 11
  - Average clustering coefficient** – 0.59
  - Number of connected components - 4

* “NA” for missing values
** For largest connected component

Network summary. Layout

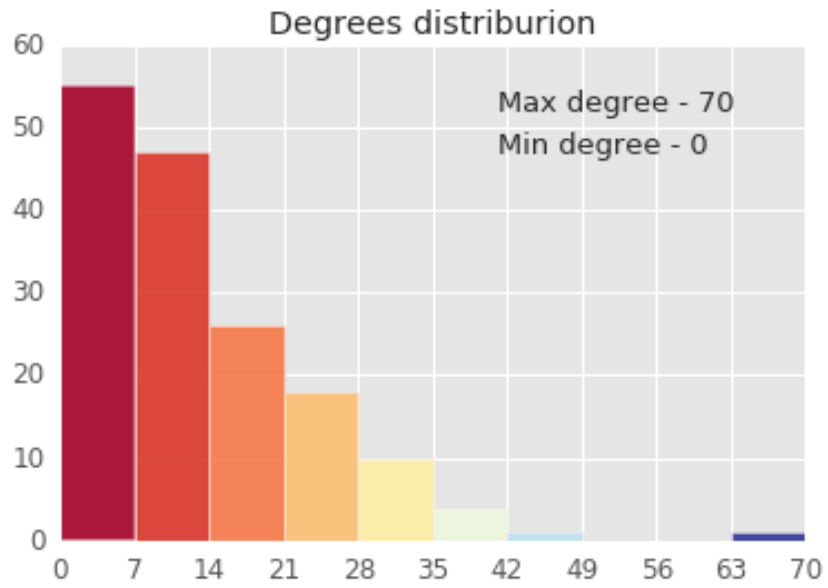


Crucial observations:

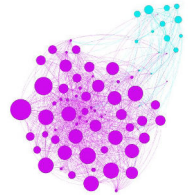
- 1 giant connected component
- 2 separate nodes – deleted accounts
- 3 distinct communities
  - hometown
  - 2 universities
- 1 node is connected with ≈50% of other nodes

Network summary. Node degrees

- Power-law nature
- The majority of nodes have <14 connections
  - Min degree - 0
  - Max degree – 70

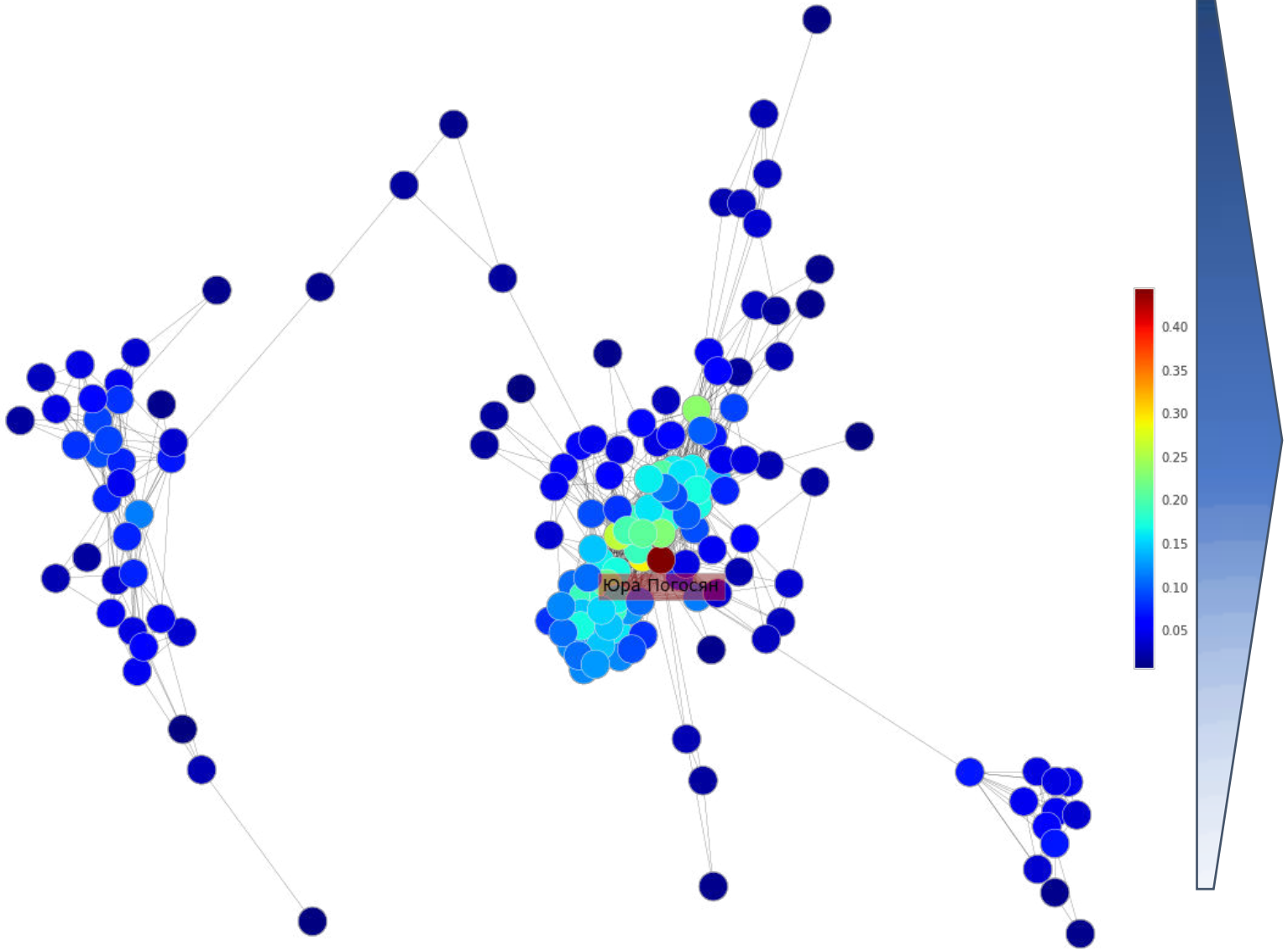


Outline



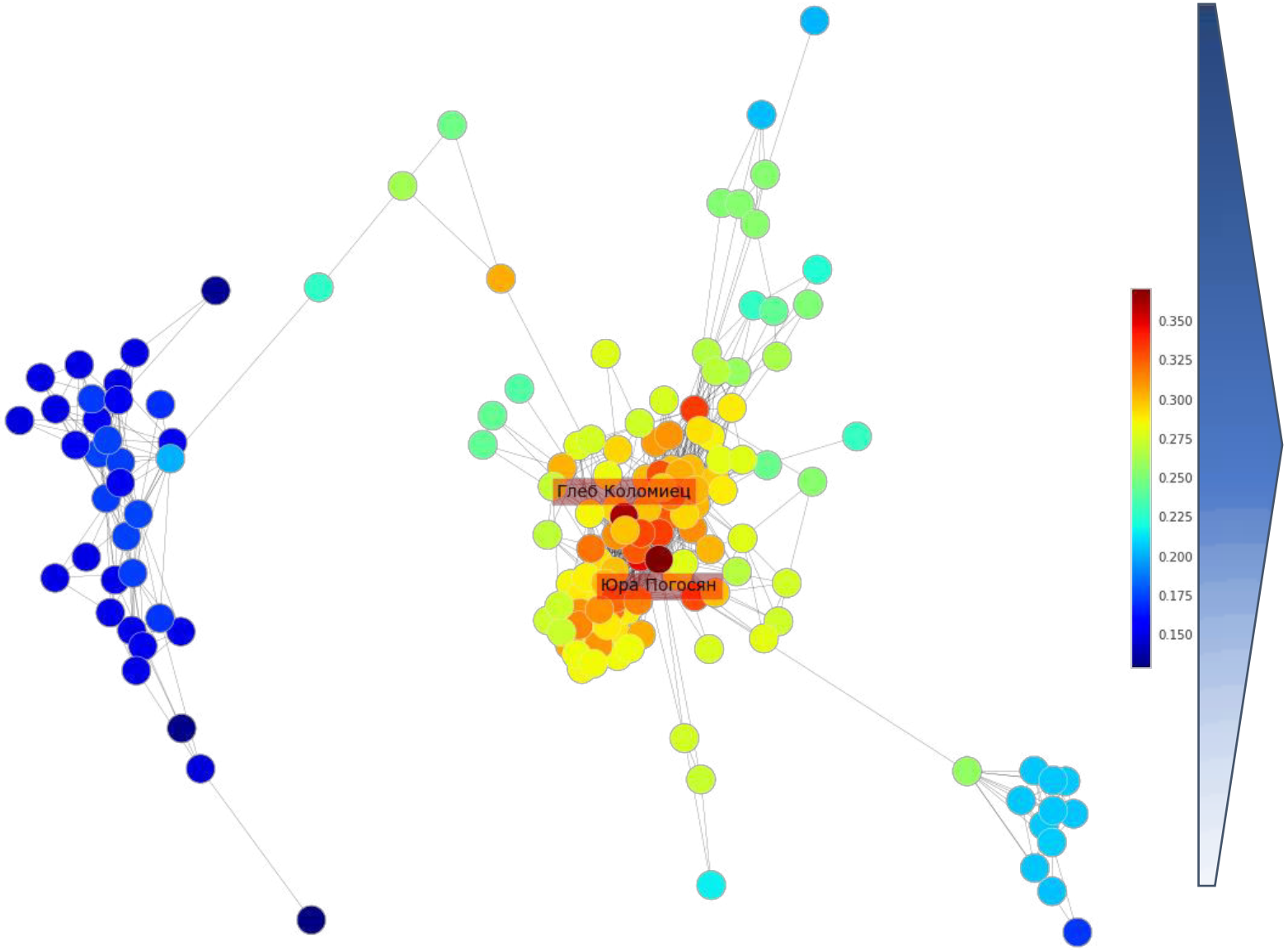
1. Network summary
2. Structural analysis
   - Centrality
   - Assortativity
   - Similarity
   - Approximating random graph
3. Community detection
   - Cliques
   - K-shells
   - Communities

Structural analysis. Degree centrality



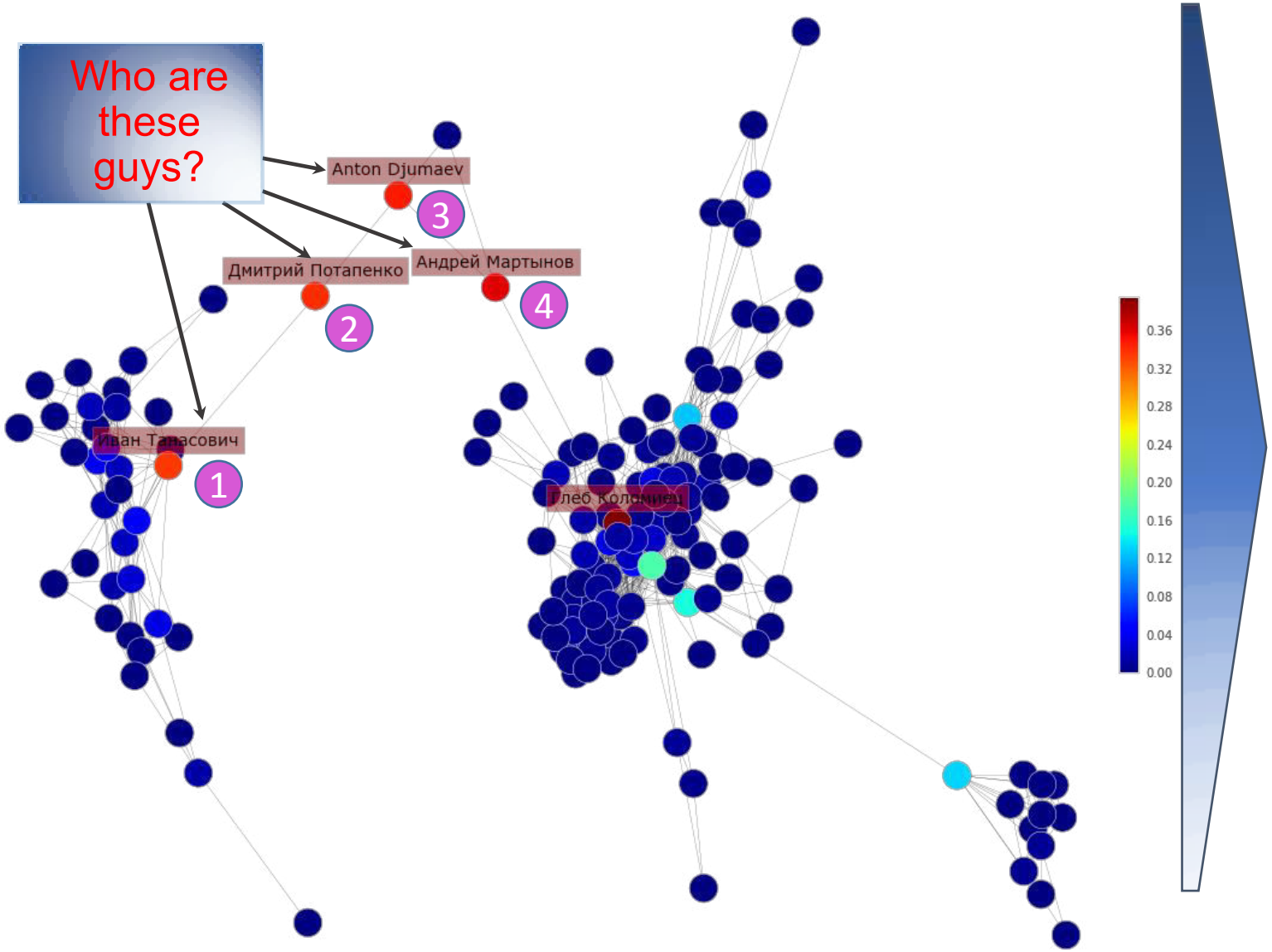
Top node **“ЮРа”** has value 0.45, i.e. knows almost 1/2 of my friends

Structural analysis. Closeness centrality



- 1st top node is the same
- 2nd top node is head-hunter “Глеб”
- More uniform distribution of metric’s values
- Nodes in giant CC have higher values – they are closer to many nodes in the graph

Structural analysis. Betweenness

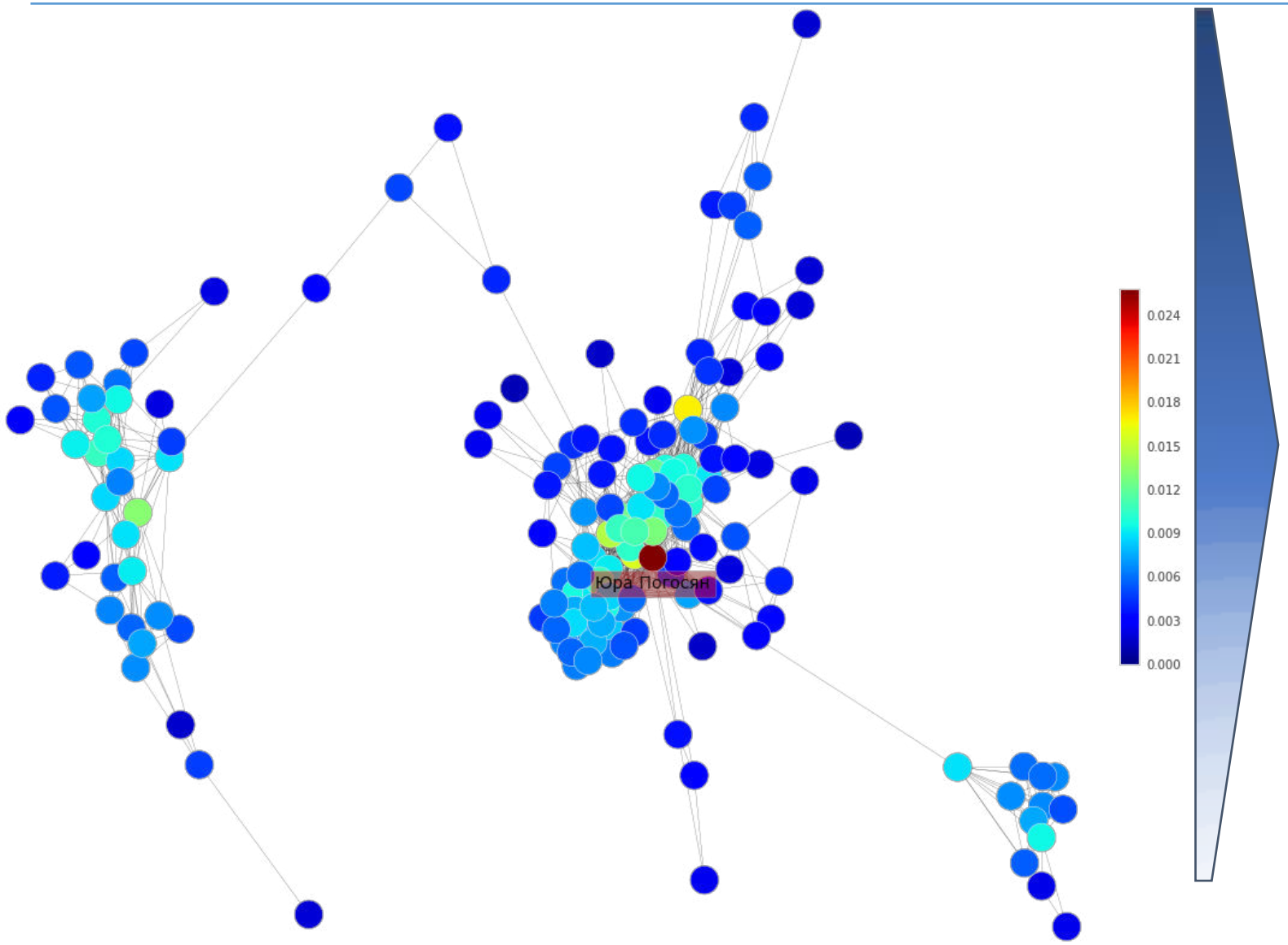


- Top 1st node is head-hunter “Глеб” – lies on many short paths

2 - well-known Russian businessman “Дмитрий Потапенко”

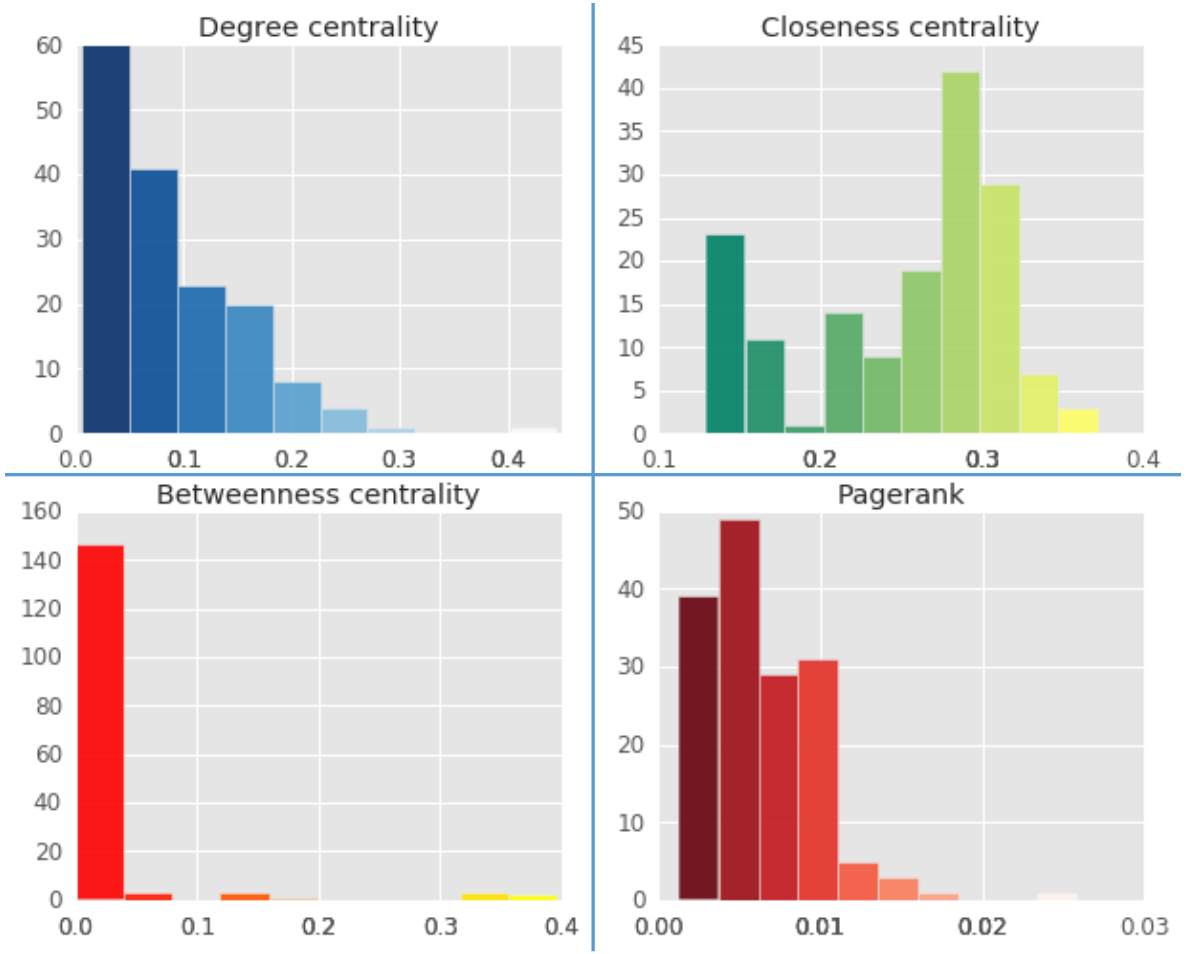
1 3 4 - the guys who just know “Дмитрий Потапенко”

Structural analysis. Pagerank



- Ranks are generally the same as degree centrality
  - same 1st top node
- Distributed more uniformly
  - nodes in small clusters still can have significant values

Structural analysis. Centrality measures

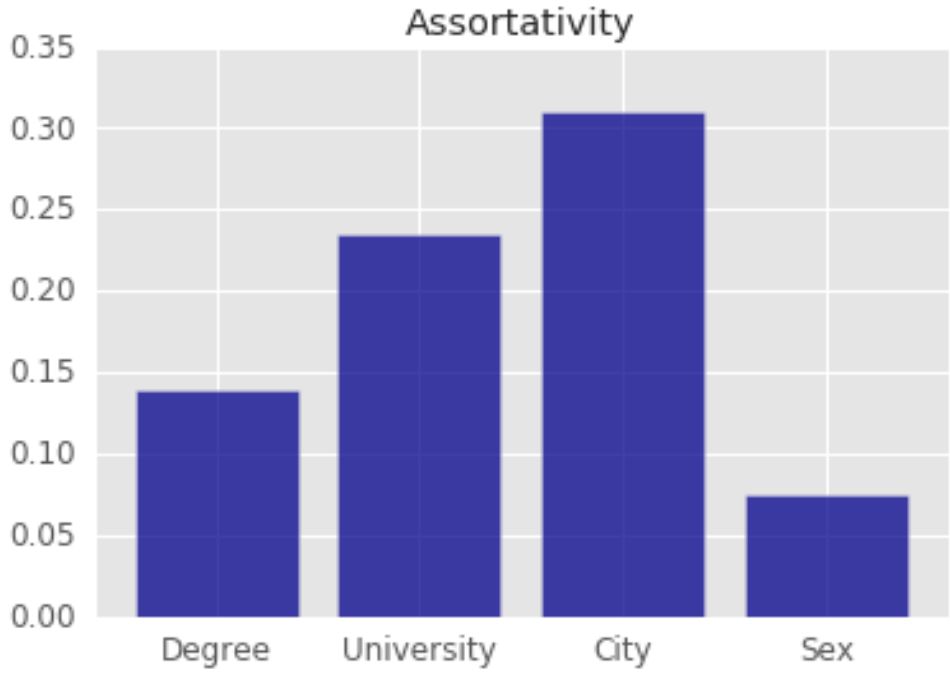


Very different distributions:

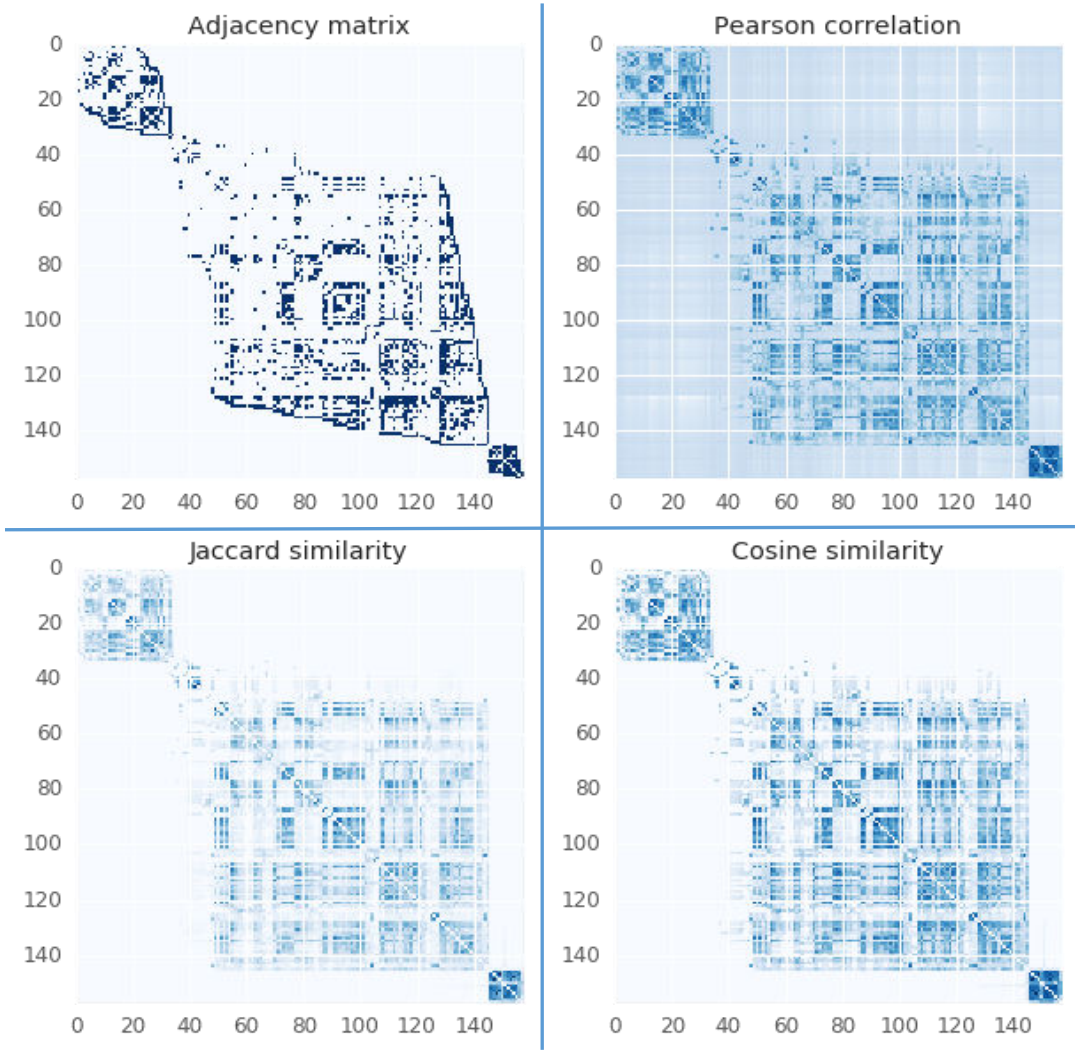
- degree centrality:
  - identical to just degree distribution
  - power-law nature
- closeness:
  - Pareto-like within communities
  - more uniform
- betweenness:
  - only 5 nodes have significant values
  - underlines those nodes which connect communities into single CC
- Pagerank:
  - almost uniform for majority of nodes
  - can be high within small almost isolated communities

Network is assortative

- City and University – the most influential attributes
- Node degree is less significant and uncertain
  - **Example:** “Дмитрий Потапенко” must have high negative degree assortativity instead of high positive
- Sex is not important



Similarity metrics reveal 3 clusters

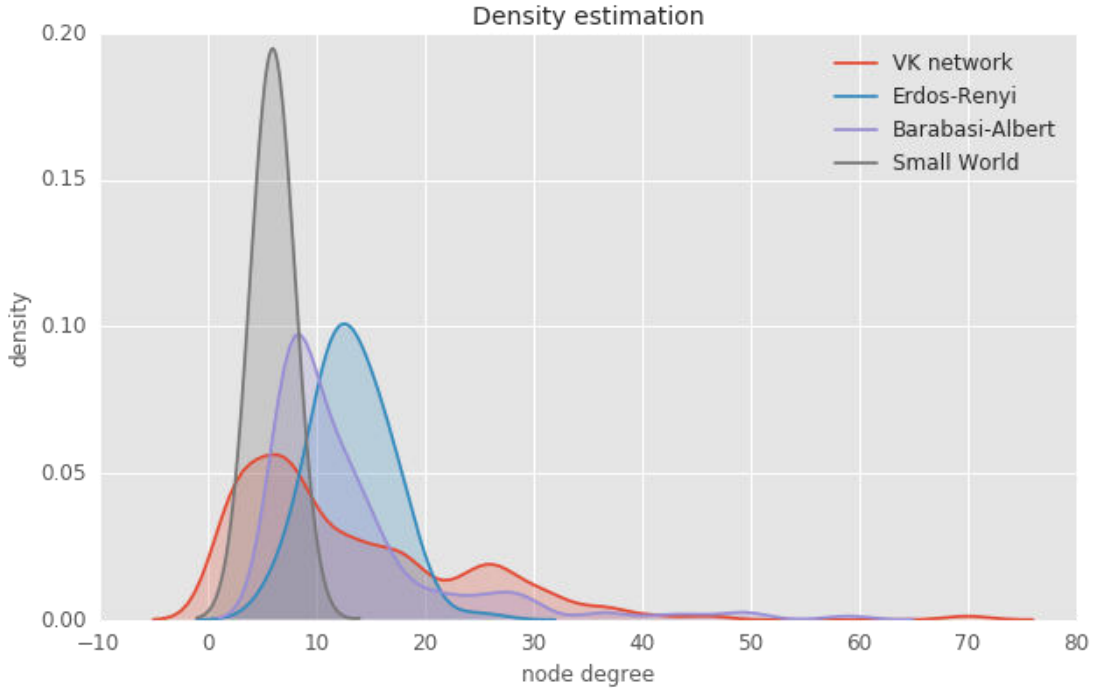


- Cuthill–McKee algorithm to rearrange adjacency matrix
- Visual output is identical
  - clear presence of 3 communities
- Order of top similar nodes – same for all metrics

Random models poorly approximate VK network

Assumptions :

- 3 metrics to compare similarity
  - density
  - diameter
  - transitivity
- same number of nodes and edges
- probability parameter in Small World is optimized*
- Degree number for BA model – average degree for VK network

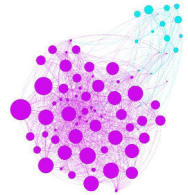


Network model	Aver clustering	Diameter
Empirical VK	0.59	11
Erdos-Renyi	0.08	4
Barabasi-Albert	0.17	3
Small world	0.57	12

- BA model has the most similar degree distribution (power-law)
  - **But:** low transitivity and small diameter
- Closest metrics in Small world model – because of optimized parameter,
  - **But:** Pareto-like density

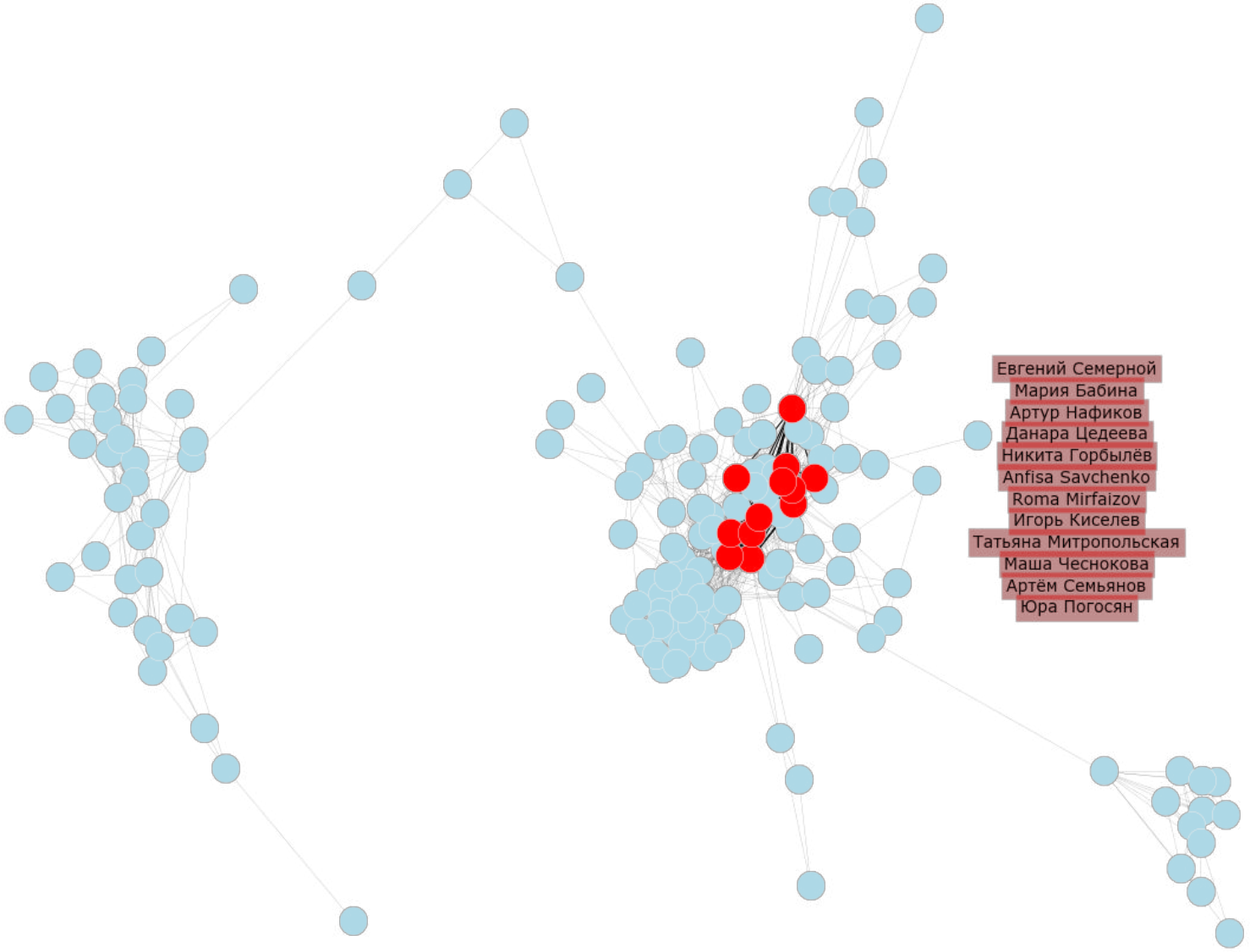
* Optimization as minimization of sum of relative differences between modularity/clustering coefficient in random model and empirical network

Outline



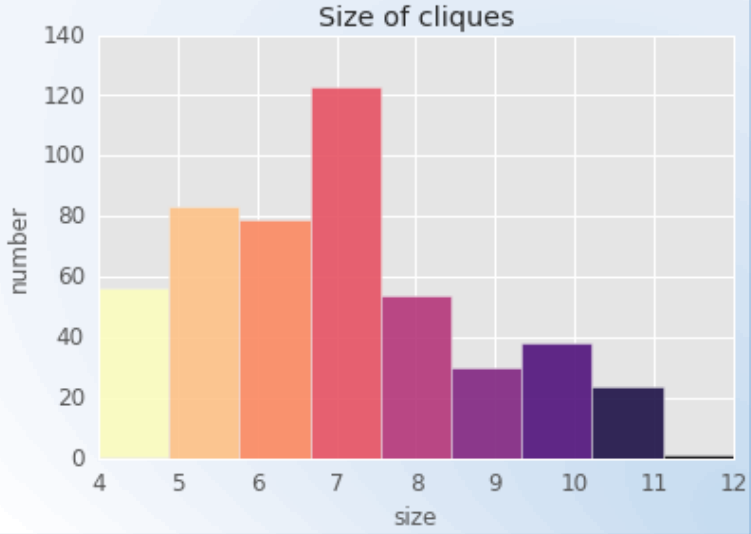
1. Network summary
2. Structural analysis
   - Centrality
   - Assortativity
   - Similarity
   - Approximating random graph
3. Community detection
   - Cliques
   - K-shells
   - Communities

Largest clique - people on the same floor of HSE dorm

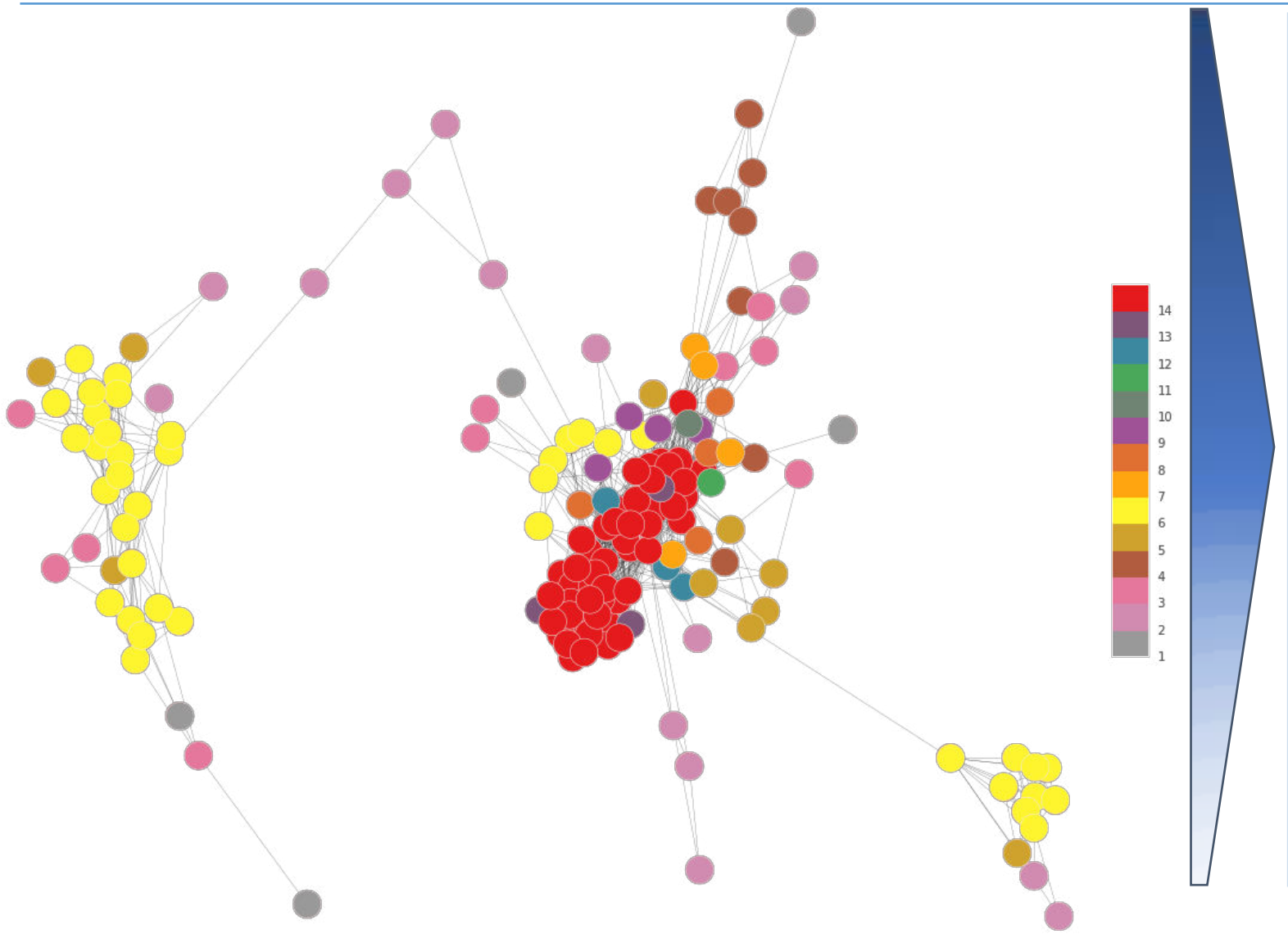


- Maximum clique size – 12
  - all lived on the 5th floor of the HSE dorm
  - located within large community

- Many cliques of large sizes:

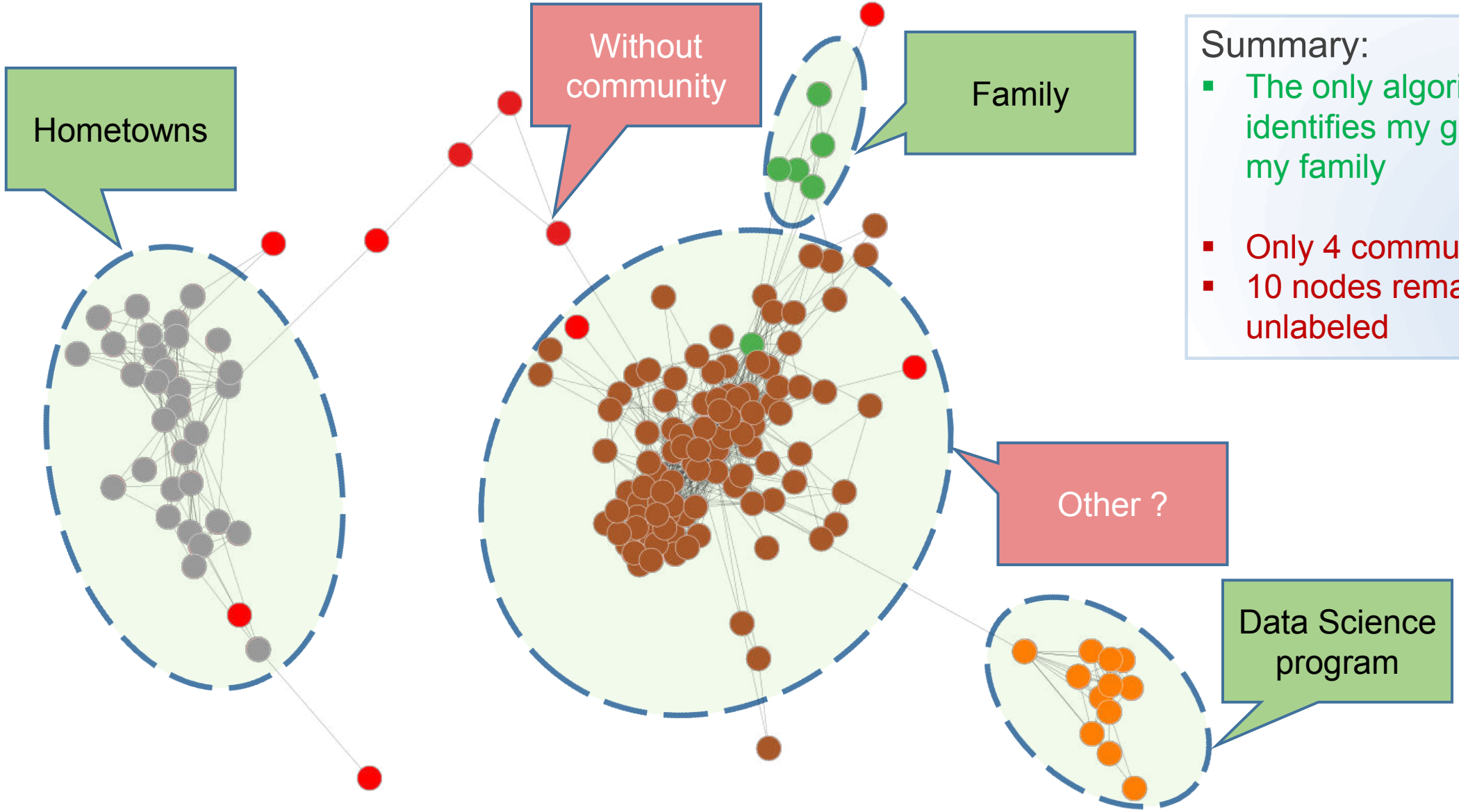


People from HSE comprise maximal 14th - shell



- Starting with K = 8, all K-shells are inside large central cluster
- 51 people in 14th – shell
  - All studied together at HSE and/or lived at the same dorm

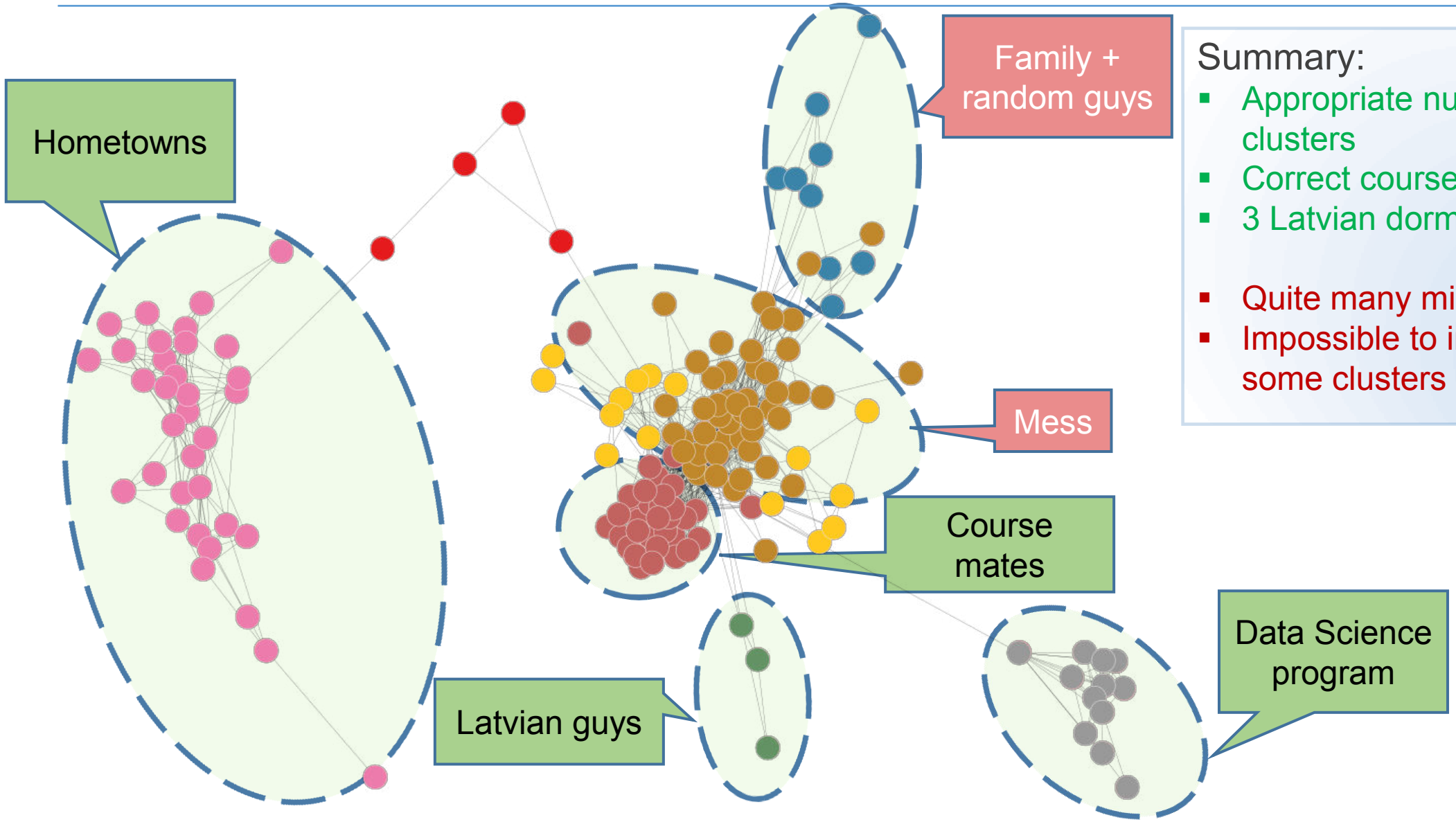
K-clique percolation finds only large obvious communities



Summary:

- The only algorithm which identifies my girlfriend as my family
- Only 4 communities found
- 10 nodes remained unlabeled

Louvain algorithm has the highest modularity but not the best output

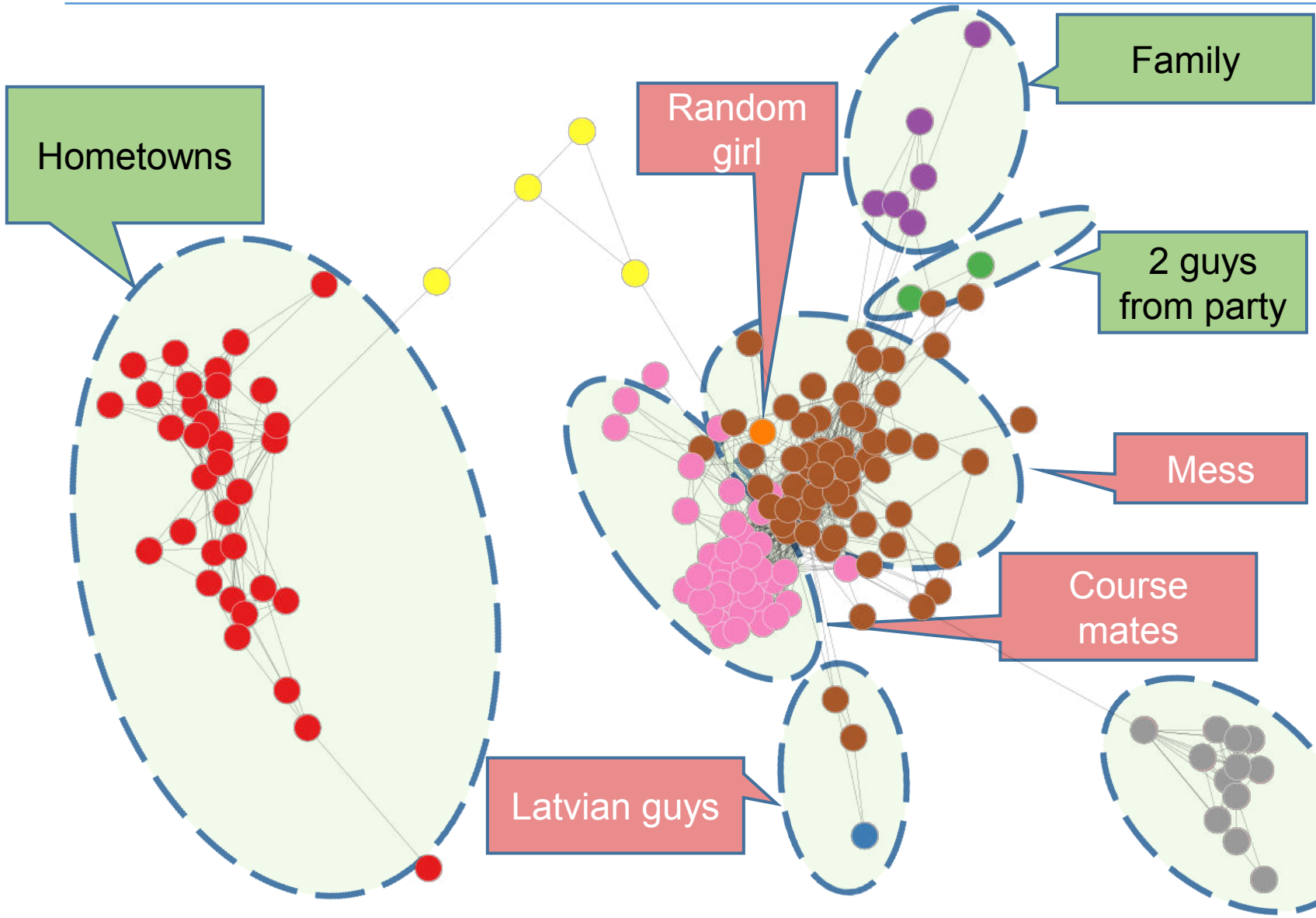


Summary:

- Appropriate number of clusters
- Correct course mates
- 3 Latvian dorm guys

- Quite many mistakes
- Impossible to interpret some clusters

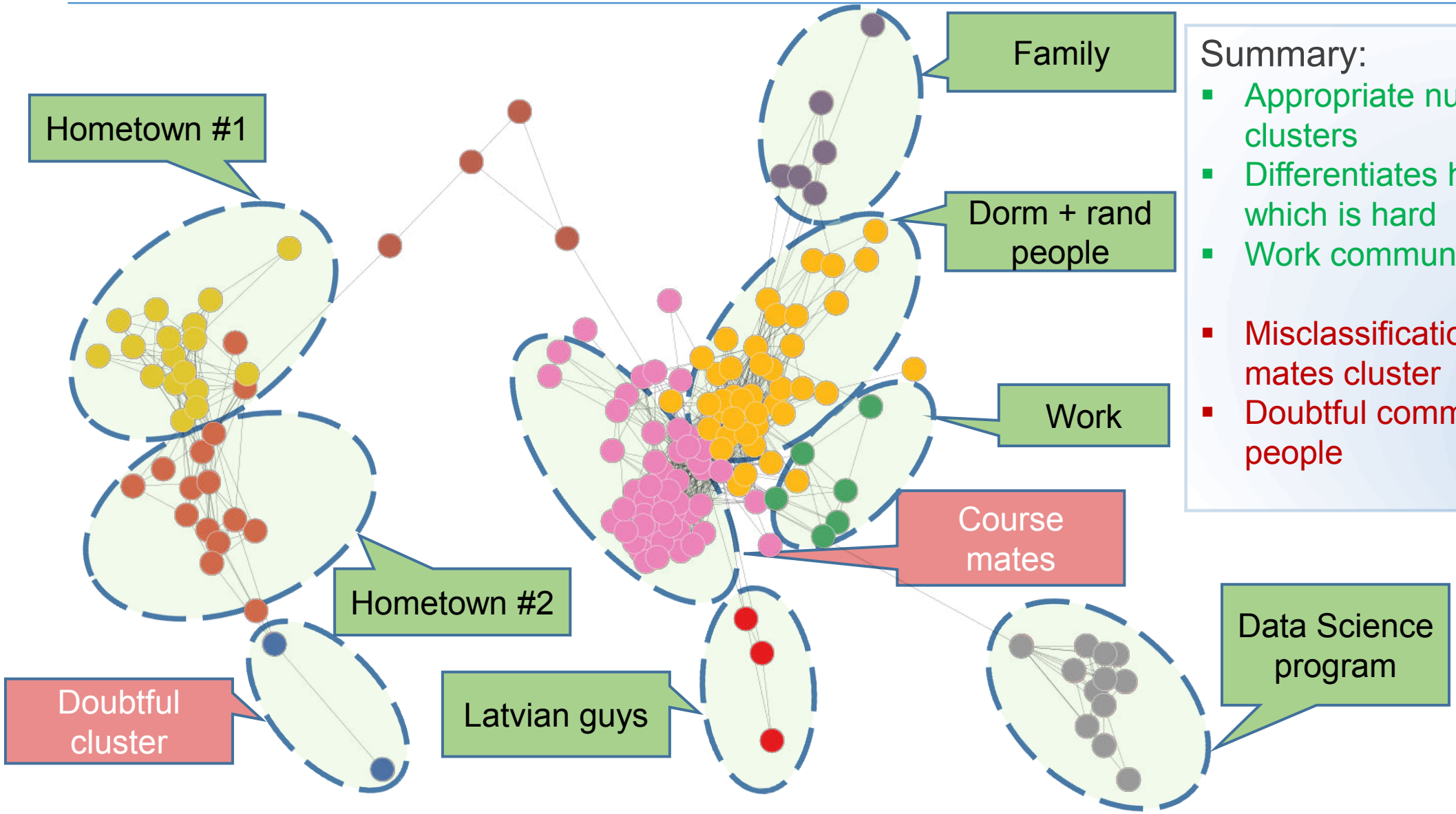
Spectral modularity optimization is worse than Louvain algorithm



Summary:

- Appropriate number of clusters
- 2 guys from party I met once
- Random girl who forms separate cluster
- Many misclassifications

Walktrap algorithm has the best performance

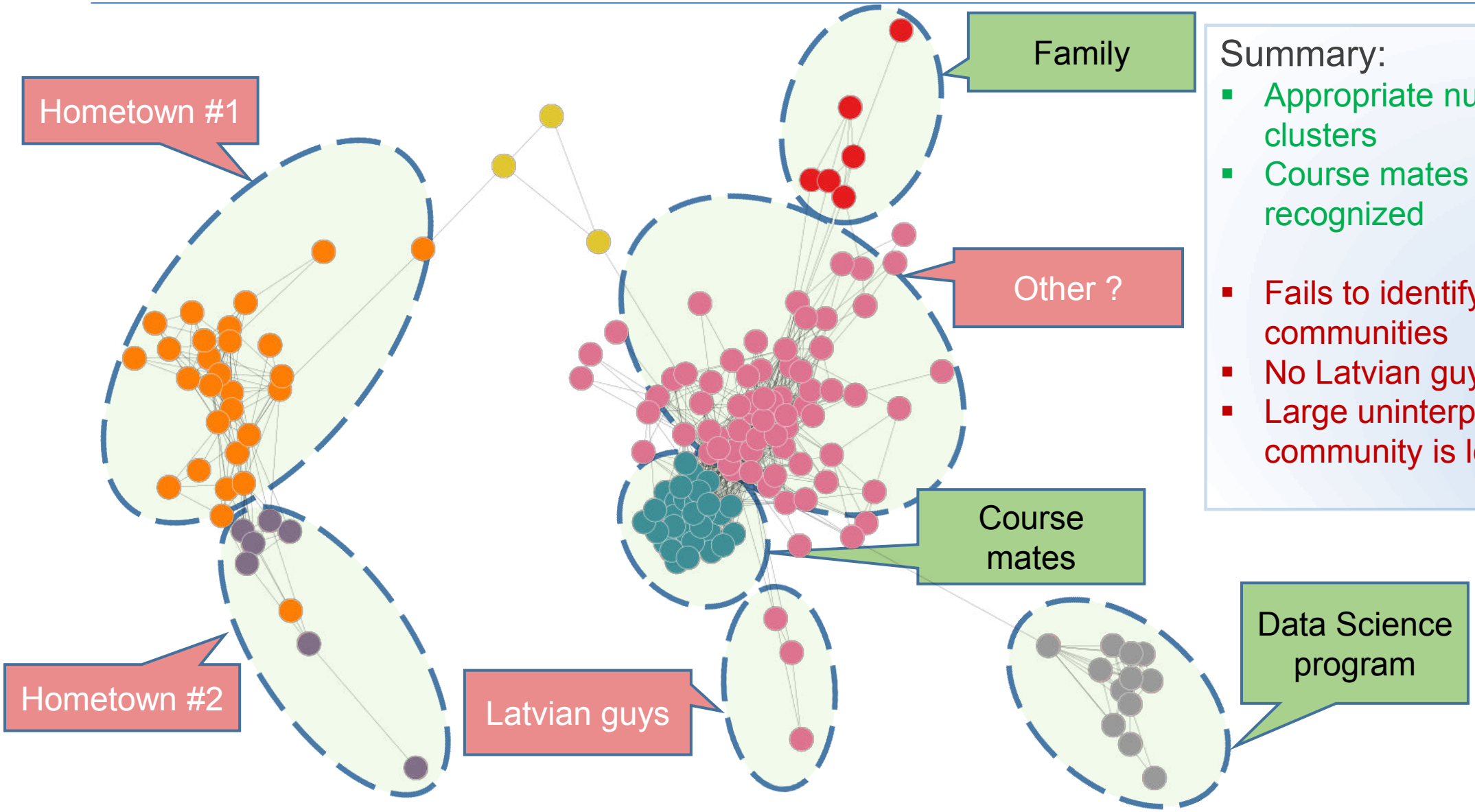


Summary:

- Appropriate number of clusters
- Differentiates hometowns, which is hard
- Work community

- Misclassifications in course mates cluster
- Doubtful community of 2 people

Label propagation algorithm is among outsiders

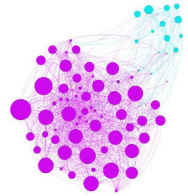


Summary:

- Appropriate number of clusters
- Course mates are recognized

- Fails to identify 2 hometown communities
- No Latvian guys
- Large uninterpretable community is left

Walktrap is the best interpretable algorithm



Highest modularity does not imply best interpretation:

Method	Modularity	Number of communities	Interpretation, rank
Louvain	0.51	8	2
Spectral modularity	0.48	9	3
Walktrap	0.46	10	1
Markov clustering	0.31	10	6
K-clique percolation	0.31	5	5
Label propagation	0.48	7	4