

CLUSTER ANALYSIS VS. GRAPH-BASED MACHINE LEARNING: The Battle Of The Strongest In Pattern Recognition



US Connect 2023



Kostiantyn Drach, PhD

Researcher, Institute of Science and Technology Austria, Klosterneuburg, Austria/Intego Group, Kharkiv, Ukraine

Kostiantyn graduated from Kharkiv National University with a Master's degree in Mathematics. He obtained his Ph.D. in Mathematics with a focus on geometry and topology from the Institute for Low Temperature Physics in Kharkiv, Ukraine. Over the last several years, he has participated in various research projects in pure and applied mathematics. Kostiantyn is a core external member of an R&D team of data scientists studying clinical data using methods of applied topology.

Iryna Kotenko

Site Lead, Intego Group, Kharkiv, Ukraine

Iryna graduated with a Master's in Quantitative Management in Health Analytics from Duke University. Her seventeen-year career as a statistical programmer spans numerous CROs as a clinical SAS programmer and data analyst. Over the last ten years, she has held the position of Site Lead for Clinical Programming Services, where she leads a group of over 200 statistical programmers in Ukraine, Poland, and the US. She is also heavily involved in a partnership program with Kharkiv National University and University of Central Florida.

Goals of the experiment:

- Use two groups of unsupervised machine learning algorithms – **clustering** and **community detection on graphs** – to reveal hidden patterns in an experimental clinical study
pattern = a pronounced subgroup of patients with ‘similar’ in-study behavior
- Perform a comprehensive comparison of the competing algorithms in terms of quality of revealed patterns (**the battle**)
battlefield = a graph representation of the data (**topological model**)
built by methods of **Topological Data Analysis**
- Improve quality of clustering algorithms by filtration based on community detection (**better together**)
- Use statistical analysis to compare our findings with the results of the original study

Experimental study: NIDA CENIC-P1S1

Design of the study

- 839 nicotine smokers
- baseline + 6 visits
- 7 treatment groups (depending on nicotine content in investigational cigarettes)

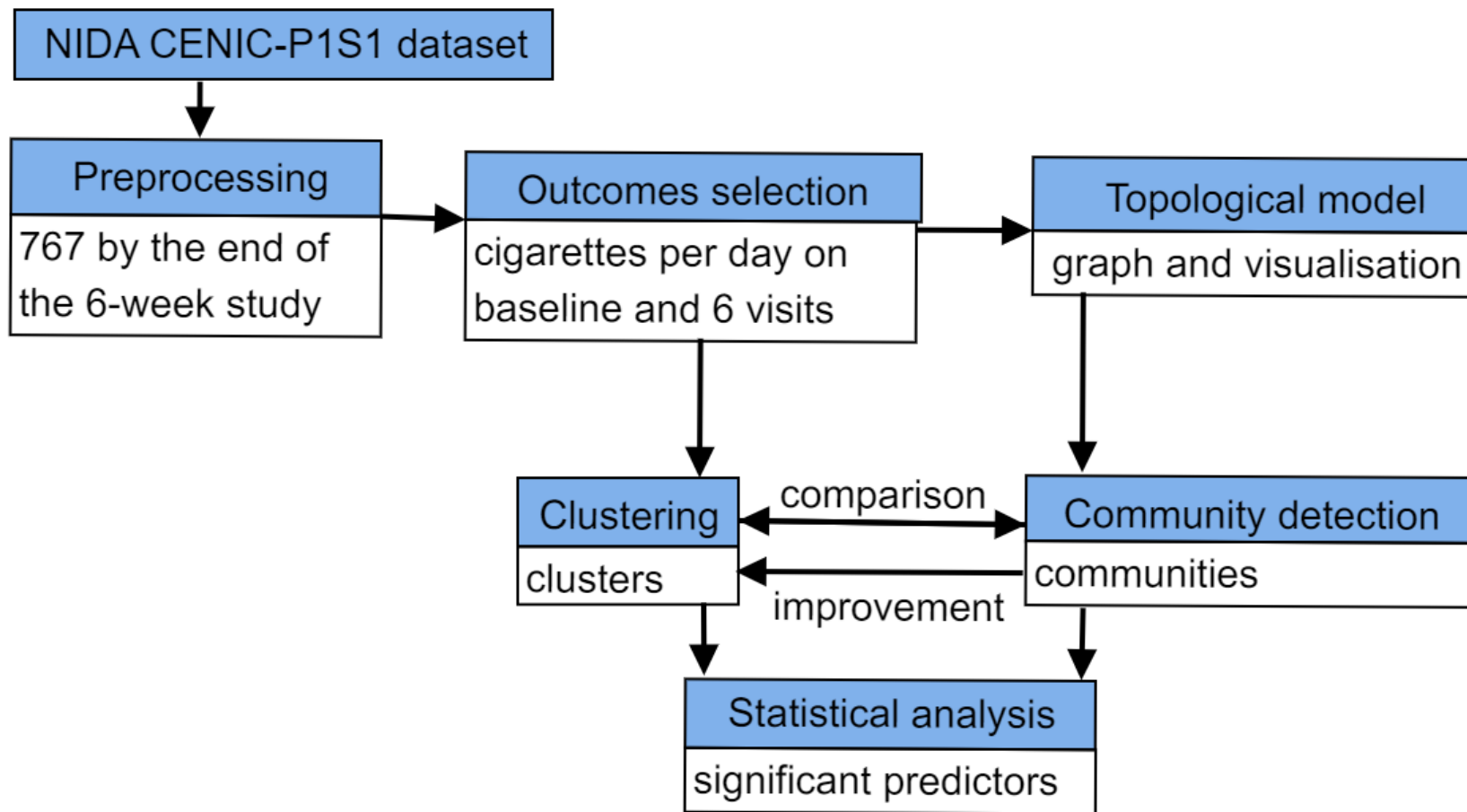
Data collected per visit

- Average number of cigarettes per day (CPD)
- Physical measurements
- Tobacco Use History
- Demographics
- Environmental and Social Influences on Tobacco Use
- Questionnaires
 - nicotine dependence, craving to smoke, intention to quit and nicotine withdrawal, depression and stress, health, perception of risk, other dependency

Outcomes

Predictors

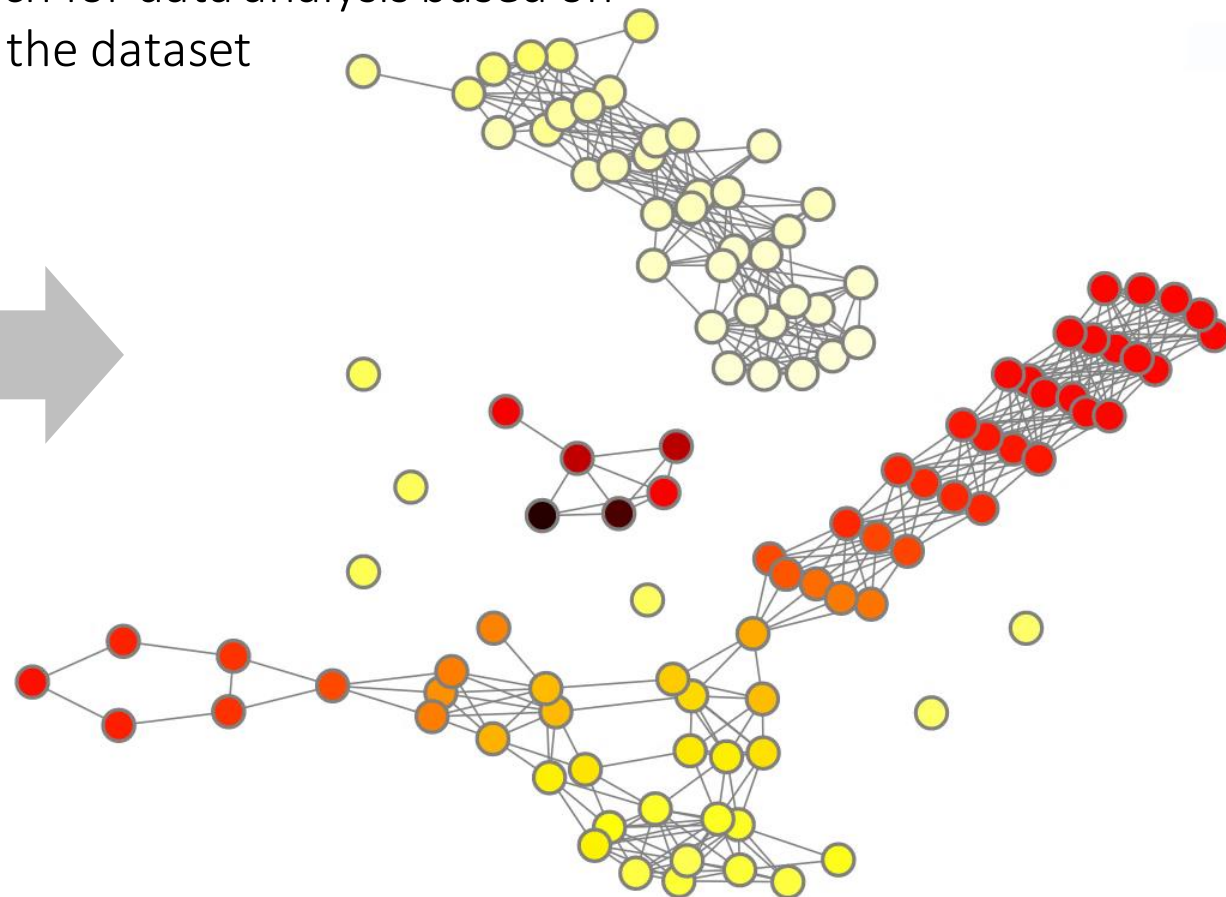
Design of the experiment



Topological Data Analysis (TDA) is a modern assumption-free approach for data analysis based on geometric properties of the dataset

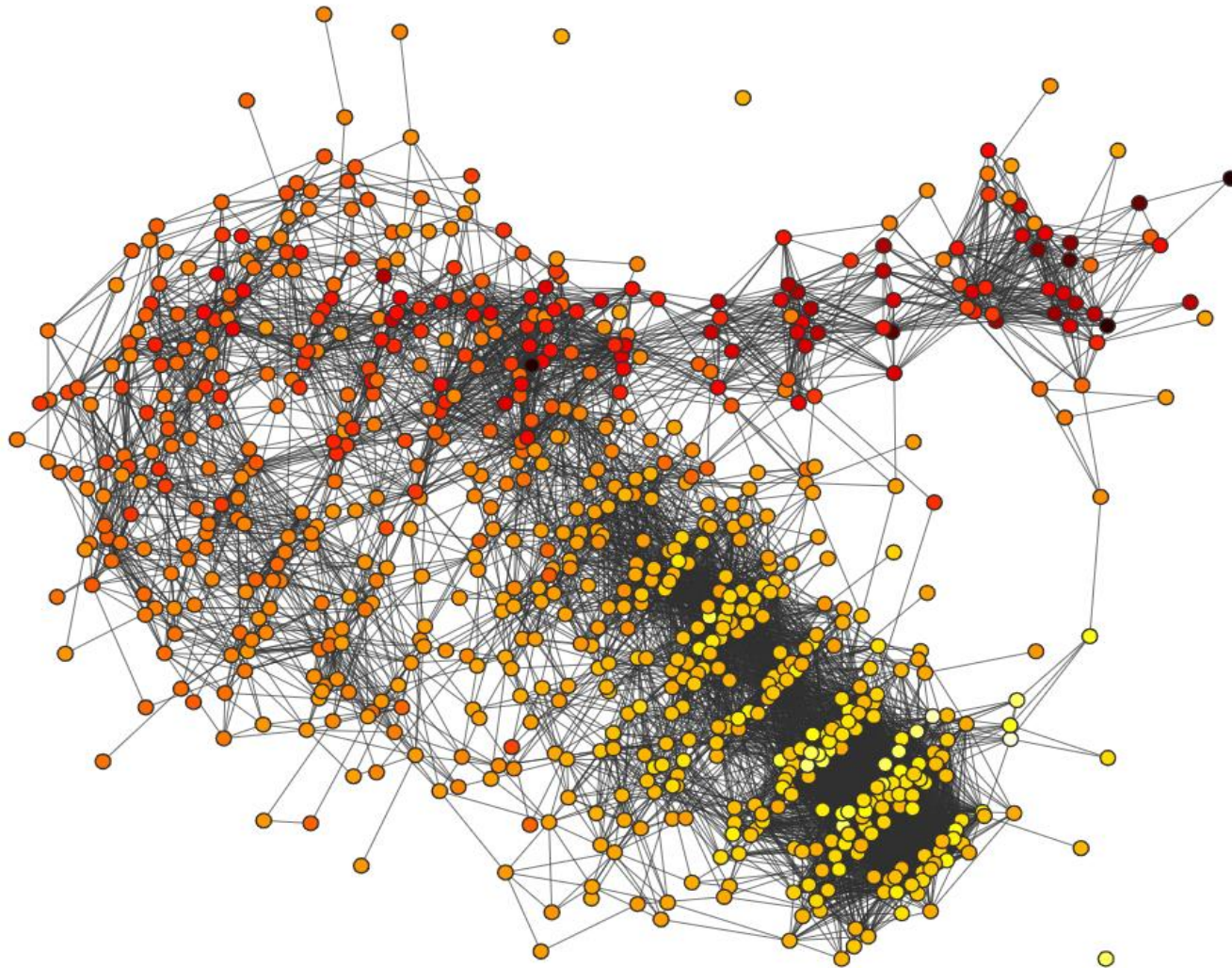


Each **node** represents one subject



Two nodes are connected if they are **similar**, within **stratifying bins**, in terms of **outcomes**

Topological model of NIDA CENIC-P1S1 dataset



smoke less  smoke more

- Each node on the **graph** corresponds to 7 weekly measures of cigarettes smoked per day (**CPD**) by a subject
- Two nodes are connected with an **edge** if the corresponding subjects have **similar** smoking behavior
- **Similarity** is measured by a combined metric based on correlation between CPD trends and on their mean values
- **Colors** of nodes reflect the deviation of the CPD mean rate from the baseline

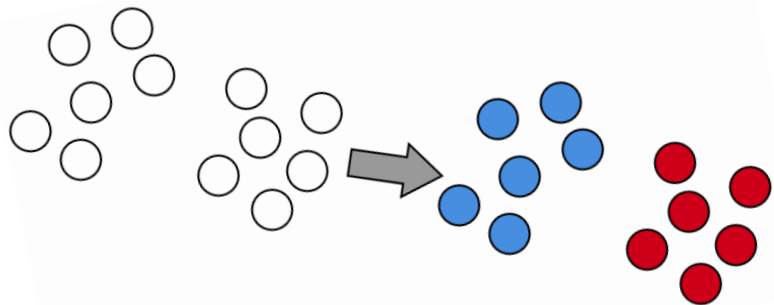
Algorithms for pattern recognition

Clustering

input: data + dissimilarity measure

k-medoids

DBSCAN

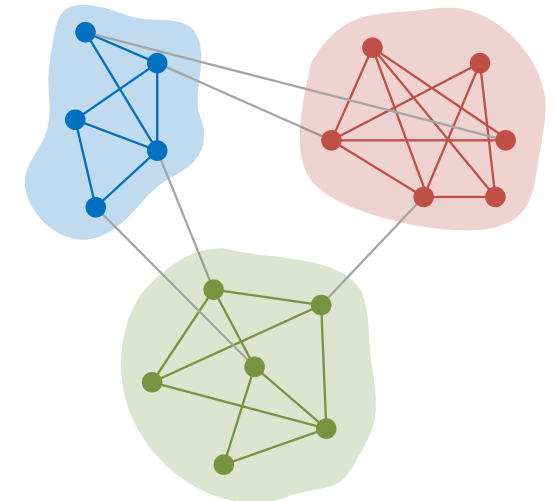


Community detection

input: topological model of the data (graph)

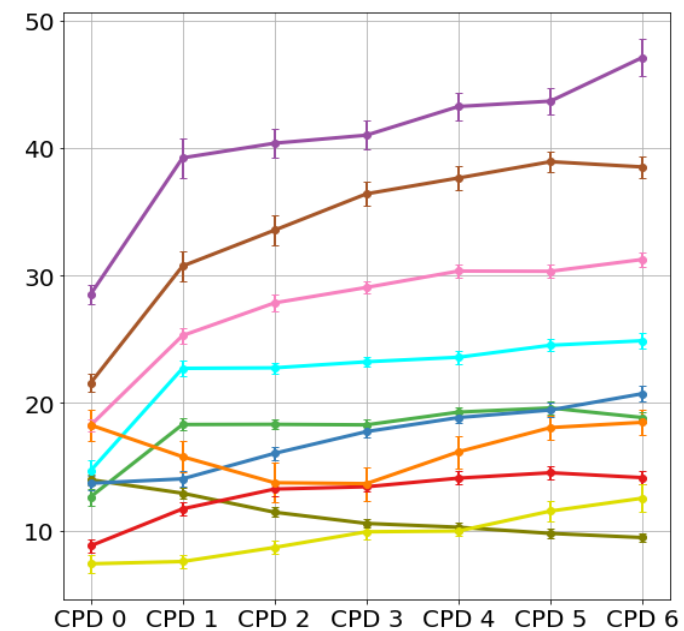
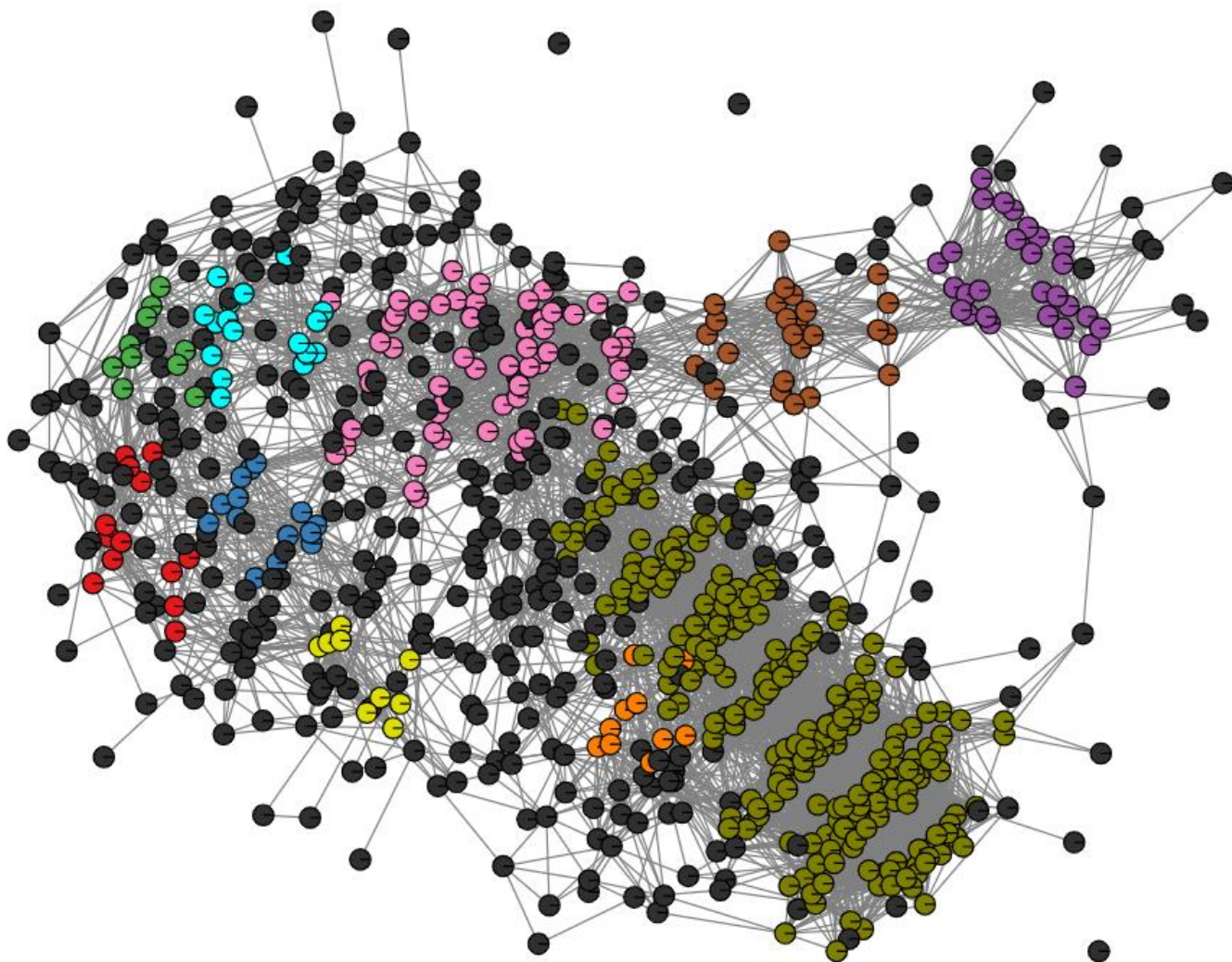
clique percolation

Girvan–Newman



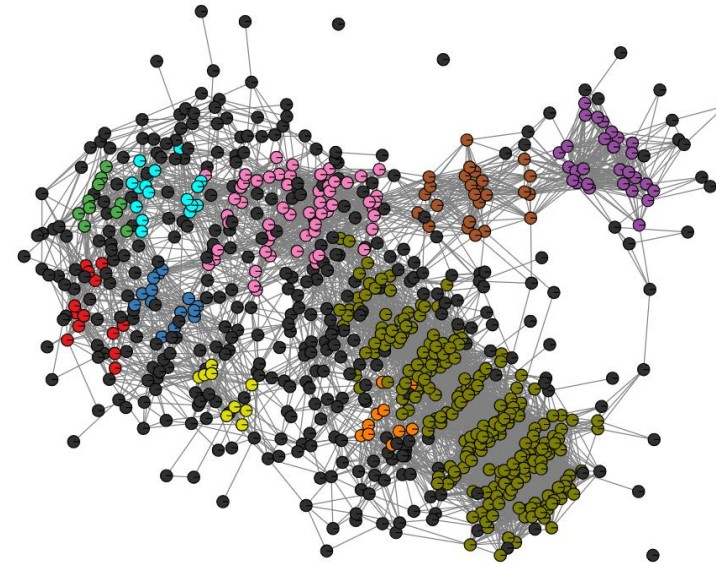
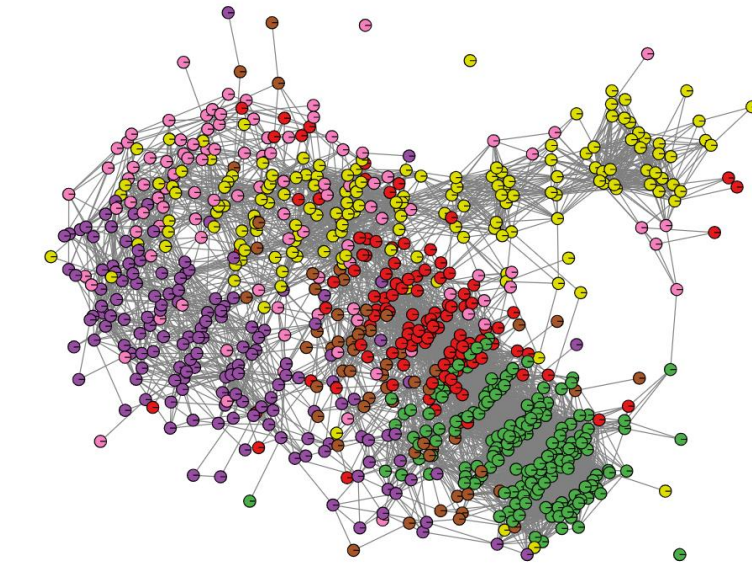
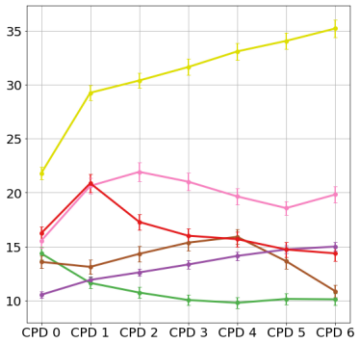
Please [download the paper](#) to get more information on clustering and community detection algorithms

Clustering and community detection results: clique percolation

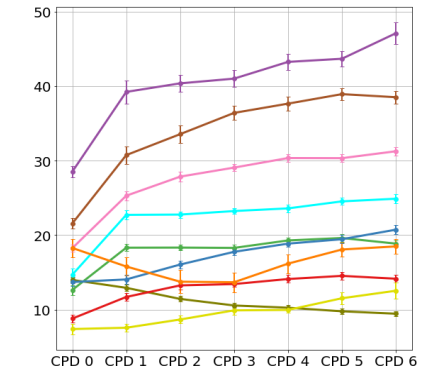


Clustering and community detection results

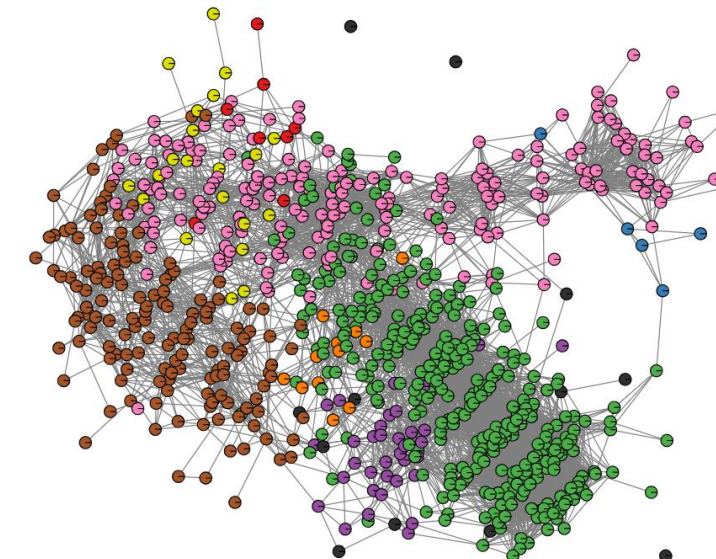
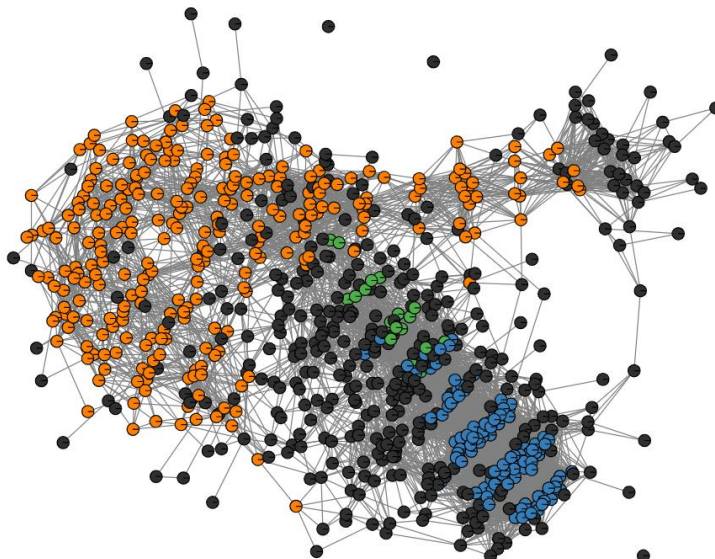
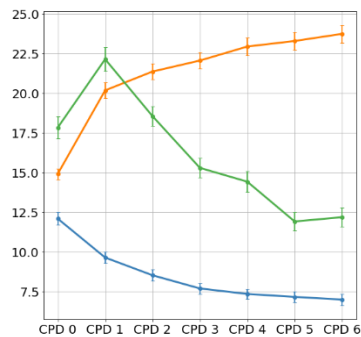
k-medoids



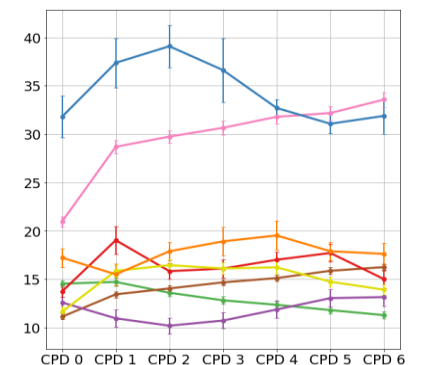
Clique percolation



DBSCAN



Girvan–Newman



Clustering and community detection: results of the battle

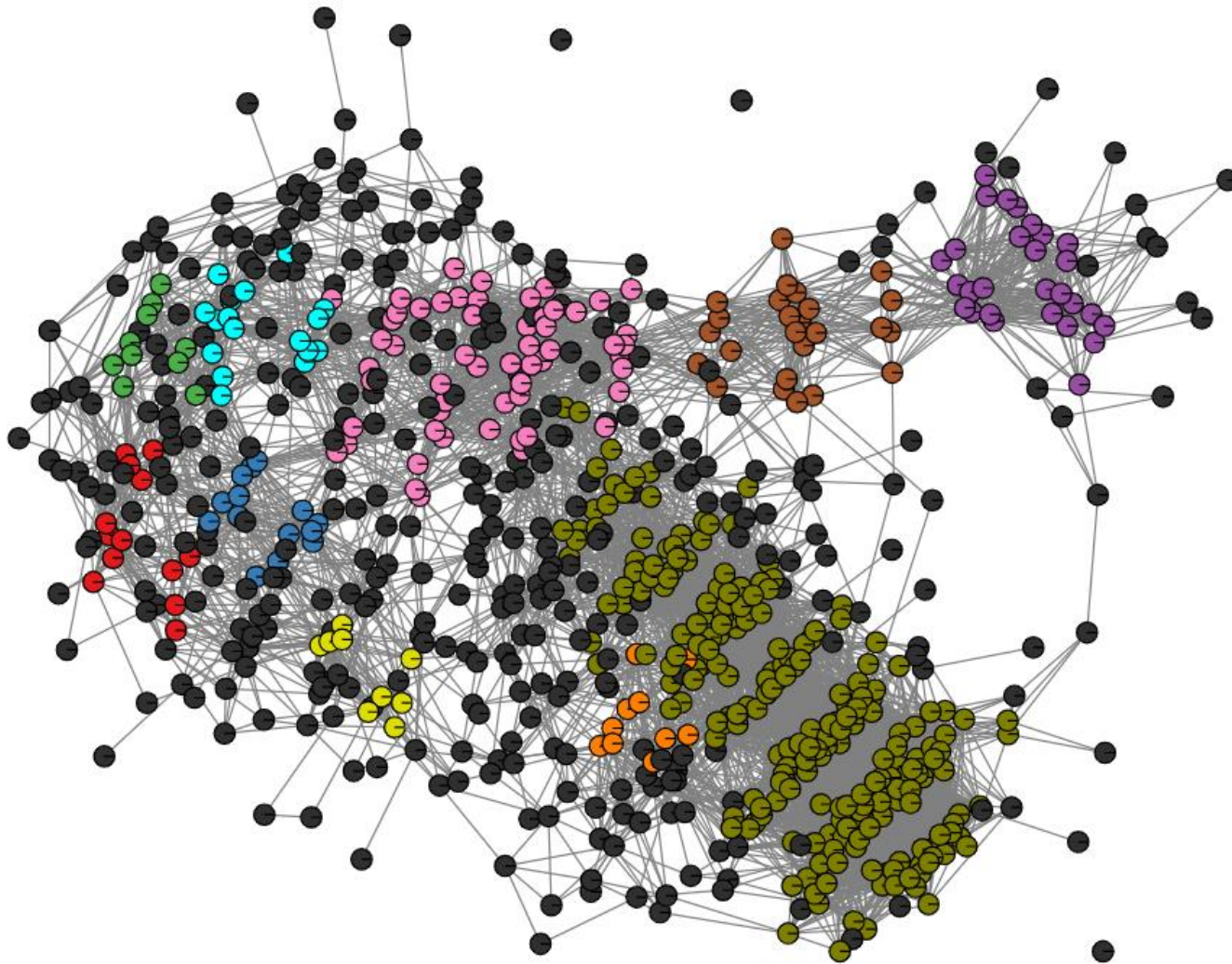
	weighted density	performance	modularity	silhouette score	DB score	DB score mean	av std	av distortion	max distortion	av outliers	av std cmdm	max distortion cmdm	av distortion cmdm
Clique percolation	0.212	0.917	0.363	0.232	1.386	0.586	9.136	9.275	65.72	0.254	0.474	0.104	0.018
Girvan-Newman	0.115	0.734	0.472	0.050	1.299	0.729	13.15	41.59	157.01	0.298	0.615	0.307	0.067
DBSCAN	0.100	0.867	0.330	0.534	0.521	0.410	12.82	57.44	158.51	0.121	0.437	0.114	0.045
k-medoids	0.125	0.837	0.529	0.243	1.258	0.817	18.04	59.43	123.86	0.191	0.634	0.127	0.067

Better together: improving quality of clustering using community detection

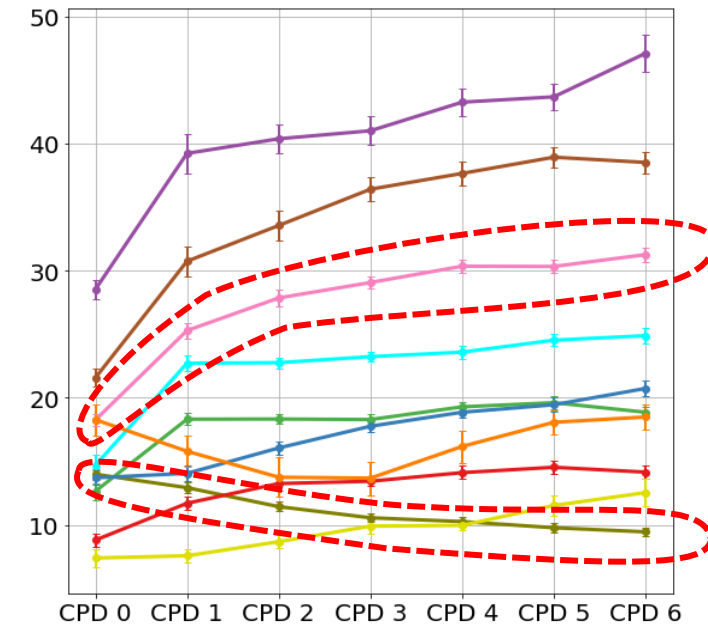
- Remove outliers given by clique percolation as an input to k -medoids clustering (**pre-** or **postfiltration**)

	weighted density	performance	modularity	silhouette score	DB score	DB score mean	av std	av distortion	max distortion	av outliers	av std cmdm	max distortion cmdm	av distortion cmdm
k-medoids	0.125	0.837	0.529	0.243	1.258	0.817	18.04	59.43	123.86	0.191	0.634	0.127	0.067
k-medoids with postfiltration	0.139	0.934	0.401	0.220	1.257	0.675	14.952	22.08	60.77	0.270	0.545	0.069	0.026
k-medoids with prefiltration	0.157	0.945	0.383	0.279	1.630	0.688	13.450	19.05	55.93	0.183	0.488	0.071	0.024

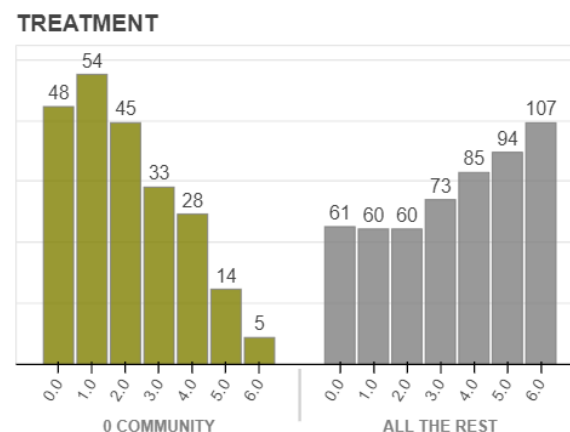
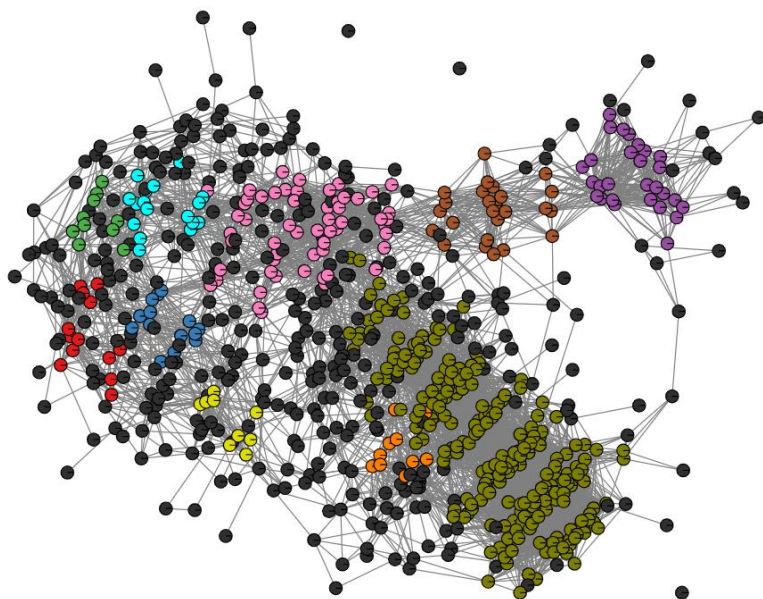
Statistical analysis of percolation communities



- 10 communities (patterns)
- 409 nodes belong to patterns
- the biggest communities:
olive (227) and pink (56)



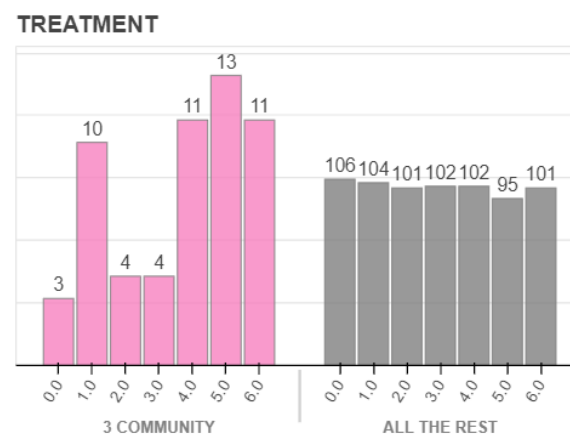
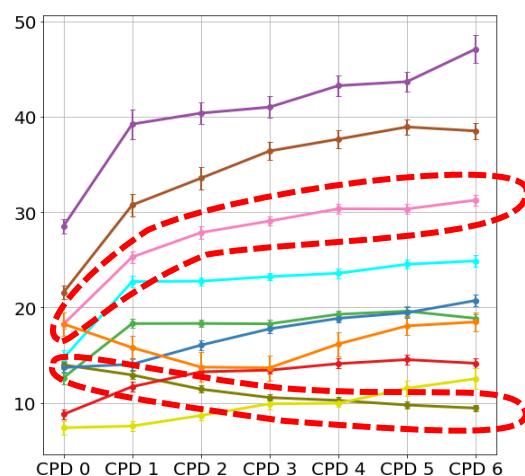
Statistical analysis of percolation communities



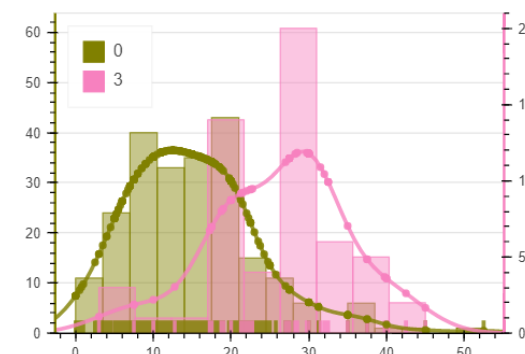
Olive community:

Significant difference in treatment groups, which confirms results of the original study

New discovery: **pink community**



'Cigarettes Purchase Task' Questionnaire



Please [download the paper](#) to learn more about statistical findings

Presentation takeaways:

- Two groups of pattern recognition techniques, namely **clustering analysis** and **community detection on graphs**, were studied on an experimental dataset
 - Most popular algorithms in each category were selected for comparison
- The underlying graph representation of data (the topological model) was built using **Topological Data Analysis**
- It was confirmed that **community detection algorithms on topological models can reveal statistically significant and more clear patterns different from those revealed by clustering**
 - Clique percolation won the battle for the best method by the majority vote of 13 quality scores
- Pre- or postfiltration made by clique percolation can improve the clustering

Please [download the paper](#) to get more information about the design of the experiment and more findings discovered



THANK YOU

Intego Group, LLC
2300 Maitland Center Pkwy, Ste. 240,
Maitland, FL 32751
Phone: +1 (407) 512-1006
www.intego-group.com/tcdm

