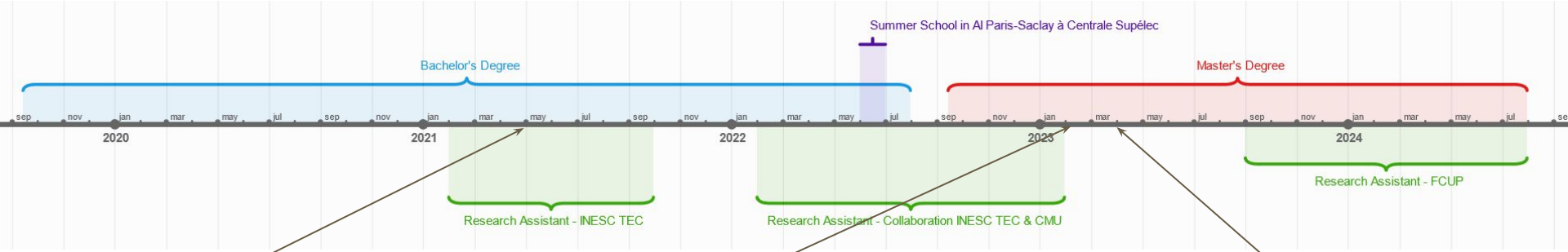
From Networks to Insights: Analyzing Graphs With and Without Machine Learning

Pedro Campos Vieira

Why Machine Learning? Why Graphs?



Artificial Intelligence: A Modern Approach by Stuart Russel and Peter Norvig

- AI as a real science and not just a label for a collection of methods.
- Possible to formalize any real-world phenomena such that computers can replicate and learn them.

Network Science by Albert-LászLó Barabási;
Network Science Course

- Graphs can represent everything. They are intuitive representations that retain a lot of formalism.
- Joining a very powerful method of viewing problems (graphs) with a very powerful method of solving problems (AI) leads to success.

Learning from Data by Yaser S. Abu-Mostafa; Vapnik–Chervonenkis theory

- The “Learning Problem” can be formalised and characterized by well-founded probabilistic and statistical principles.
- Defined traditional machine-learning cycles as theoretical backed steps.

Selected Non Graph Related Research

- Designed and implemented new spatio-temporal data structures to improve the space and time complexity of the state-of-the-art methods.
- Major applications are in geometrical rendering and tracing and information query.

To be published soon 😃

- Developed and implemented methods based on reinforcement learning and genetic algorithms to extract information from encrypted TCP connections.
- Applied methods that building on extracted information measure temporal sequence similarity.
- Major applications are in deanonymizing Dark Web Traffic of Mix Networks like Tor.

Technical report not available to the public.

- Formulated a version of t-SNE to be applied to data streams. The methods focus on ensuring space and time complexity in infinite data scenarios.
- Worked on fast methods to select points based neighborhood density to accommodate infinite data and on methods based simple tessellations to combat drift.

S+t-SNE - Bringing Dimensionality Reduction to Data Streams.10.1007/978-3-031-58553-1_8

Dynamic Mechanisms forming Networks of MOBA Matches

Motivation

- Competitive online gaming is one a popular hobbies. MOBA games such as League of Legends and Dota 2 are among the most played ones.
- Understanding what is the dynamics and topology of the network that the system behind the matchmaking creates can help with:
  - The study of how scam attempts and harassment propagates.
  - Determine the level of influence of players in the network.
- Such studies aim at developing a smoother, safer and more enjoyable experience to players.

Method

- Query 50000 matches starting from a random player in the most populous game rank and performing branch-limited breadth-first search (BFS).
- Transform the data obtained in a network with nodes being games (direct interaction between 10 players) and edges connecting matches with common players.

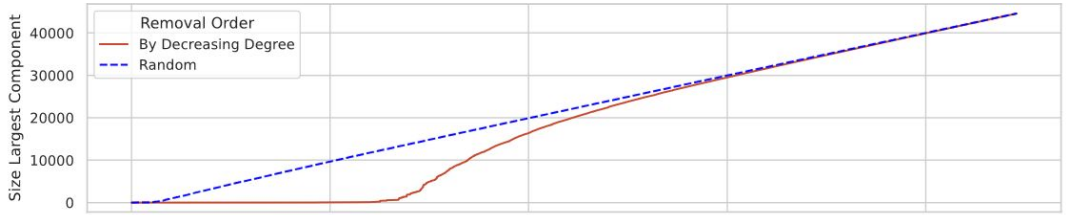
Dynamic Mechanisms forming Networks of MOBA Matches

Findings

- High percentage of repeated players, between 27.65% and 47.66%, meaning players often play in close circles. Matchmaking system cannot efficiently shuffle players. This is further amplified by different players tending to play in a different but specific times of the day.

- Bianconi-Barabasi process (10.1209/epl/i2001-00260-6) with α = 2.47 and k_{min} = 48, some initial attractiveness and link removal seems to be the best model to describe the truncated power-law the network presents. This aligns with the interference expected from the rank-based matchmaking system.

- Exposure curves and resilience tests through percolation show the expected behaviour with respect to the model estimated.
  - ⚠Key Takeaway: Attempts to mitigate scam and harassment should focus on “immunize” key players rather than try to prevent it everywhere.⚠



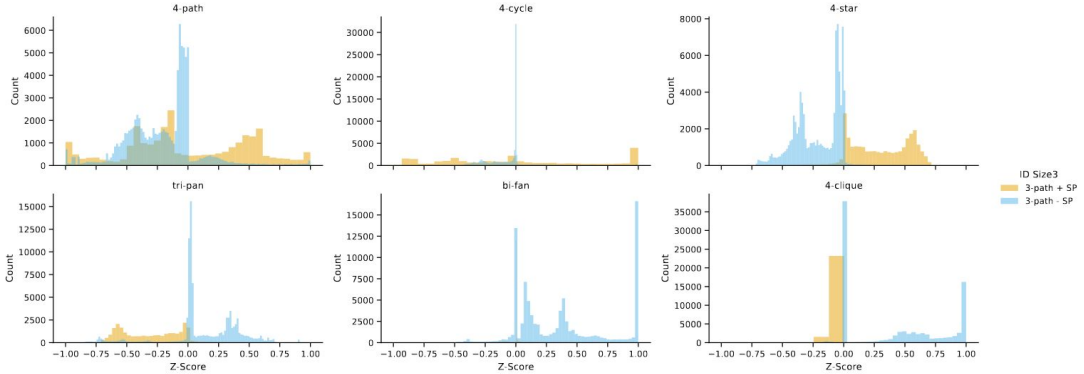
Evolution of the size of the largest component as nodes are removed of the network

Using Graph Neural Networks to Find Motifs

- Motifs as a tool e.g. Motif2vec, Motif Graph Attention Networks, GNNExplainer, TempME
  - These works do not predict motifs, they assume they already exist or create them with traditional methods.
- Counting occurrences of graphs e.g. Simple MPNNs, Subgraph GNNs, etc.
  - 1. Are fully bounded by the expressivity (for this task) of the models used. No workaround to avoid it.
  - 2. Suffer from high variation in the number of occurrences of the graphs of interest as the size of the graph(s) to analyse increases.
- Directly predicting motifs e.g. GROVER, MGSSL, MICRO-Graph, MotifFiesta, SPMiner
  - 3. The application of a NULL model is either nonexistent or unsatisfactory.
  - 4. Lack of a score to compare the relevance of different motifs found.
  - 5. Lack of interpretability for the mechanism used to generate the motifs.
  - 6. Difficulty to control the size of the graph(s) branded as motifs.
  - 7. Ignore everything that is not a motif.

Our Approach

- Removes ambiguity by training a specific generic set of graphs, S, that is known to be important for network analysis. Allows to easily get a complete description of S. ← limitations 5, 6 and 7.
- Train the model to return relevant results by integrating the null model in the target variable. ← 3 and 5.
  - Break theoretical results regarding expressivity of models for motif finding. ← 1.
- Normalise the score used as target. ← 2 and 4.
  - The normalisation forces an algebraic dependence between graphs with the same number of nodes.
  - Constraining the scores between -1 and 1 allows comparison between networks of different sizes.
  - Removes instability of the variance between predicted scores on networks of different sizes*.
- Pick S such that multi-target regression with mechanisms such as weight sharing can help the model have a strong inductive bias towards meaningful patterns**. ← Bonus



Organization of motif scores color coded based on the score of the 3-path graph

* Could be achieved with normalisation techniques e.g. few-shot, invariant risk minimization, invariant feature learning. However, it would be more complex and the two points above would be lost.

** Regression Chains should further enhance the model's capacity.

Comparison with other approaches

Graph Type	single	multi	single	multi	single	multi	single	multi	single	multi	single	multi	single	multi	single	multi
100%	2.921	2.868	1.231	0.991	1.193	0.816	1.498	0.861	1.056	1.268	2.275	2.388	1.814	2.193	2.940	2.621
		-1.814%		-19.496%		-31.601%		-42.523%		+20.076%		+4.967%		+20.893%		-10.850%
95%	0.279	0.250	0.316	0.396	0.309	0.320	0.350	0.347	0.547	0.502	0.517	0.489	1.294	0.741	1.463	0.808
		-10.394%		+25.316%		+3.560%		-0.857%		-8.227%		-5.416%		-42.736%		-44.771%
75%	0.078	0.091	0.095	0.043	0.047	0.041	0.083	0.053	0.067	0.038	0.082	0.100	0.210	0.042	0.357	0.048
		+16.667%		-54.737%		-12.766%		-36.145%		-43.284%		+21.951%		-80.000%		-86.555%
50%	0.004	0.009	0.017	0.011	0.004	0.006	0.008	0.007	0.005	0.003	0.007	0.012	0.051	0.007	0.042	0.007
		+125.000%		-35.294%		+50.000%		-12.500%		-40.000%		+71.429%		-86.275%		-83.333%
25%	0.000	0.001	0.001	0.000	0.000	0.000	0.001	0.001	0.001	0.000	0.001	0.002	0.005	0.000	0.012	0.003
		+100.000%		-100.000%		0.00%		0.00%		-100.000%		+100.000%		-100.000%		-75.000%

Multiple scores at once (multi) against a single one (single) - % variation between squared error of multi and single.

Graph Type	Count	SP	Count	SP	Count	SP	Count	SP	Count	SP	Count	SP	Count	SP	Count	SP
75%	0.469	0.222	0.377	0.323	0.326	0.199	0.322	0.227	0.361	0.173	0.680	0.298	0.509	0.172	0.625	0.278
		-52.584%		-14.378%		-38.799%		-29.618%		-52.117%		-56.225%		-66.154%		-55.554%
50%	0.339	0.116	0.236	0.109	0.175	0.056	0.174	0.083	0.214	0.083	0.311	0.126	0.368	0.072	0.358	0.079
		-65.907%		-53.718%		-68.007%		-52.079%		-61.143%		-59.456%		-80.448%		-78.013%
25%	0.041	0.044	0.115	0.042	0.106	0.036	0.065	0.031	0.063	0.020	0.144	0.041	0.328	0.027	0.300	0.021
		+5.426 %		-63.960%		-66.396%		-52.797%		-68.212%		-71.614%		-91.903%		-92.887%

Motif scores directly (SP) against counting structures (Count) - % variation between absolute error of SP and Count.

- Multi-target regression: big win
- Direct estimation: benefits outweigh the prejudices. It is specially good in the first 3 graphs.
  - We expect this difference to increase in favor of direct estimation in out-of-distribution estimation.
- ⚠ Key Takeaway: Our approach should benefit any model with dimensionality smaller than the size of the largest graph in S (assuming some relation between graphs of S). ⚠

Research Questions I am interested to explore (*very* open to other stuff in similar areas)*

- How does the characteristics of a network/system affect its capability to learn? Can we derive any insights using topological features? You et al. 2020, Papamarkou et al. 2024
- How can geometric machine learning approaches be utilized to explore key structural and functional concepts in complex networks, such as connectivity, modularity, and centrality, to gain deeper insights across domains like neuroscience? Can general knowledge systems also benefit? Luo et al. 2024, C. Vieira et al. 2024
- How do we draw more power from Graph Neural Networks? Should we pursue new designs? What current design choices limit their effectiveness? Morris et al. 2023, Müller et al. 2023, Zhang et al. 2024
- How can we artificially generate networks that accurately mimic the real-world at multiple levels? Barabasi 2016, Du et al. 2024, Kovács and Jlidi 2024
    - How can we use networks to accurately represent a wide range of real-world phenomena, such as social interactions or brain activity?
- Can we modulate complex networks as means to achieve foundational learning systems? Bommasani et al. 2022, Cheng et al. 2024
- Can we modulate diffusion and cascade behaviour as a blueprint to improve the learning process and/or learning capacity? Leskovec et al. 2007, Kipf and Welling 2017, You et al. 2020

* With a focus on Geometric Deep Learning, or more generally, Network Science and Machine Learning.

Research Questions I am interested to explore (*very* open to other stuff in similar areas)*

- How does the characteristics of a network/system affect its capability to learn? Can we derive any insights using topological features? You et al. 2020, Papamarkou et al. 2024
- How can geometric machine learning approaches be utilized to explore key structural and functional concepts in complex networks, such as connectivity, modularity, and centrality, to gain deeper insights across domains like neuroscience? Can general knowledge systems also benefit? Luo et al. 2024, C. Vieira et al. 2024
- How do we draw more power from Graph Neural Networks? Should we pursue new designs? What current design choices limit their effectiveness? Morris et al. 2023, Müller et al. 2023, Zhang et al. 2024
- How can we artificially generate networks that accurately mimic the real-world at multiple levels? Barabasi 2016, Du et al. 2024, Kovács and Jlidi 2024
  - How can we use networks to accurately represent a wide range of real-world phenomena, such as social interactions or brain activity?
- Can we modulate complex networks as means to achieve foundational learning systems? Bommasani et al. 2022, Cheng et al. 2024
- Can we modulate diffusion and cascade behaviour as a blueprint to improve the learning process and/or learning capacity? Leskovec et al. 2007, Kipf and Welling 2017, You et al. 2020

* With a focus on Geometric Deep Learning, or more generally, Network Science and Machine Learning.

Thank you.