What is a good project?

COMP 4880 / 8880

Lexing Xie

Plan of this lecture

A project, what?

From an interesting topic to a project proposal

Critiquing project reports

Mechanics

Why do a project?

The main goals is to prepare you to apply state-of-the-art network analysis tools and algorithms to an application.

If you are interested in research: doing well in Network Science project will also leave you well-qualified to do network science research.

Pragmatically, project experience is useful when you:

- Look for a job
- Apply for graduate school

What is a project like?

A project is a significant body of work that

- Articulates a *problem* of relevance
- Presents a systematic attempt to solve the problem
- Documents the problem + new understandings gained

By nature, a project needs to be new.
NOT a valid project: re-producing other people's results (from research papers), implement a known algorithm ...

The best place to start is to follow your interest + passion, and then move on to narrow down and define the specific problem and finding answers.

- From interest to project topic

I’m interested in …

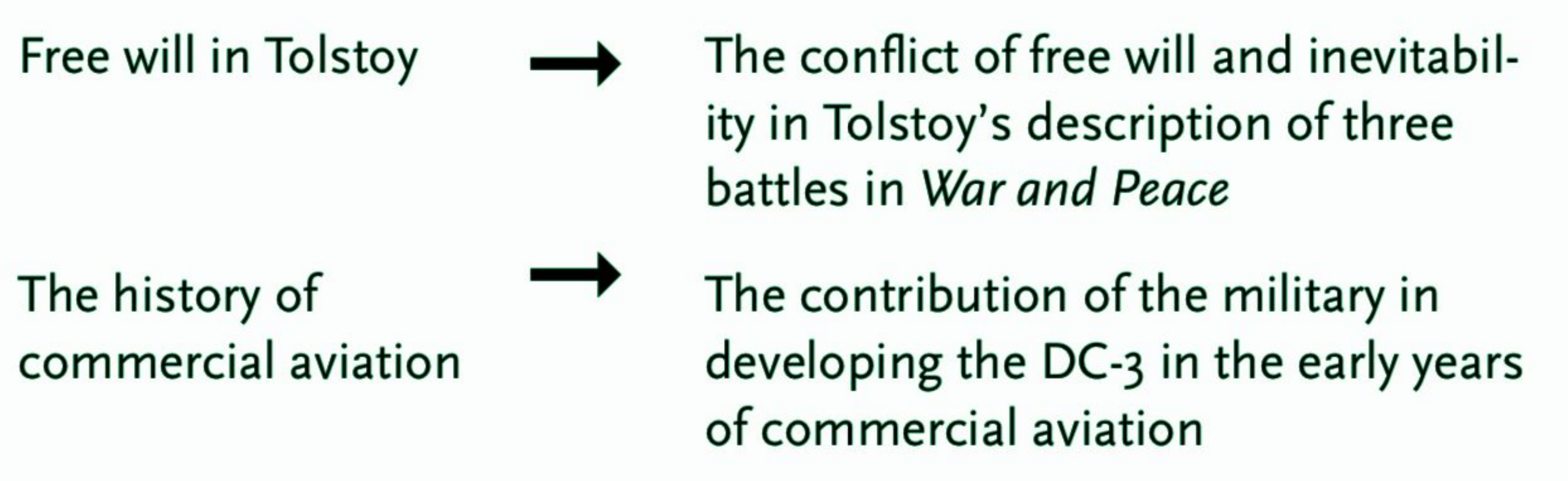
“fake news”

“electricity networks”

“neural networks in living organisms”

“how people vote”

Narrowing down a broad topic



Free will in Tolstoy → The conflict of free will and inevitability in Tolstoy's description of three battles in *War and Peace*

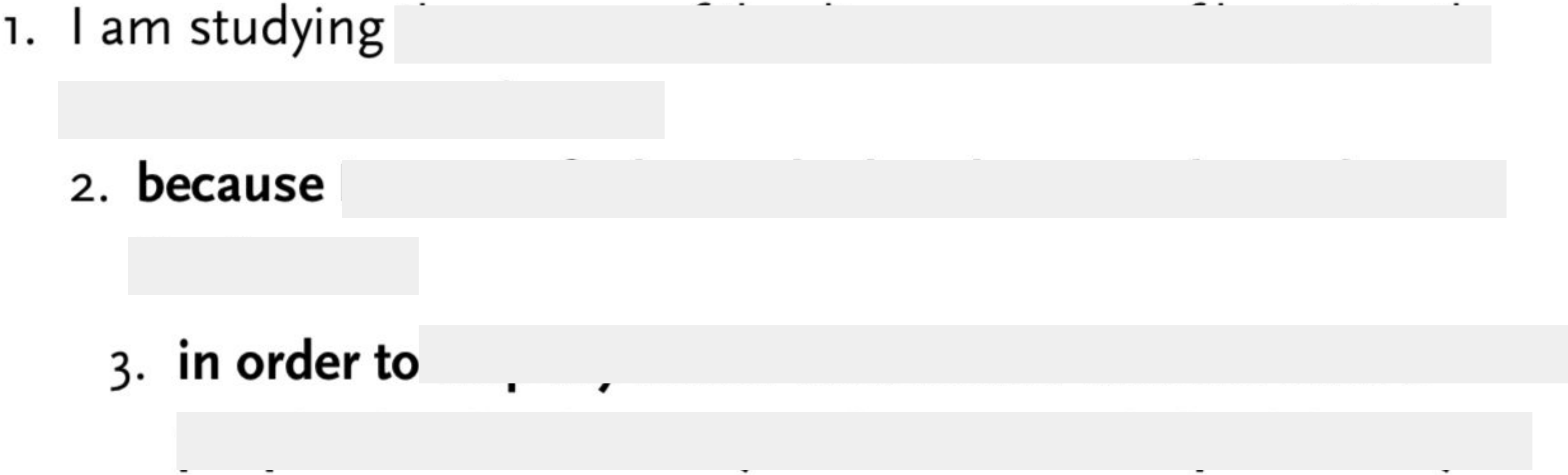
The history of commercial aviation → The contribution of the military in developing the DC-3 in the early years of commercial aviation

- Make your topic bounded (i.e. will not take a lifetime to study)
- Qualifiers help define what data to rely on for addressing the topic

Example from [Craft of Research, Booth et al]

Evaluate + articulate topic significance

Answer the following 3-part question



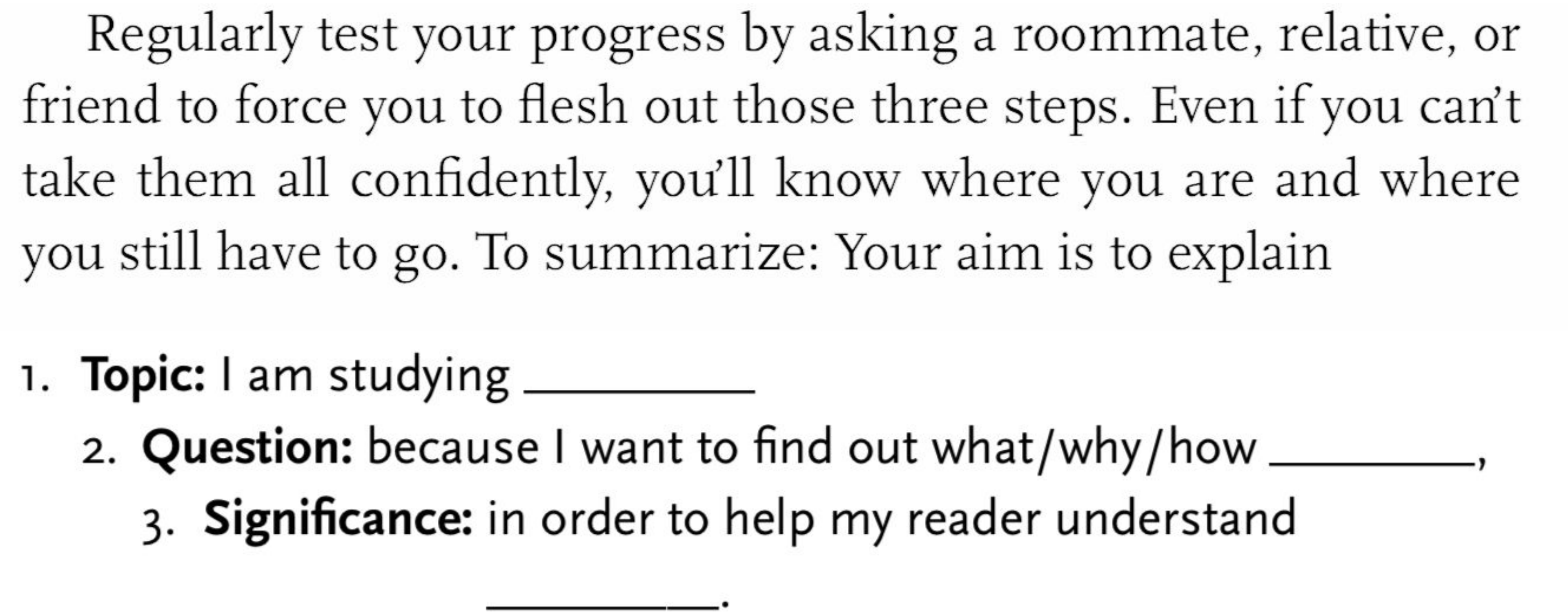
1. I am studying [redacted]

2. **because** [redacted]

3. **in order to** [redacted]

Example from [Craft of Research, Booth et al]

Test for topic significance

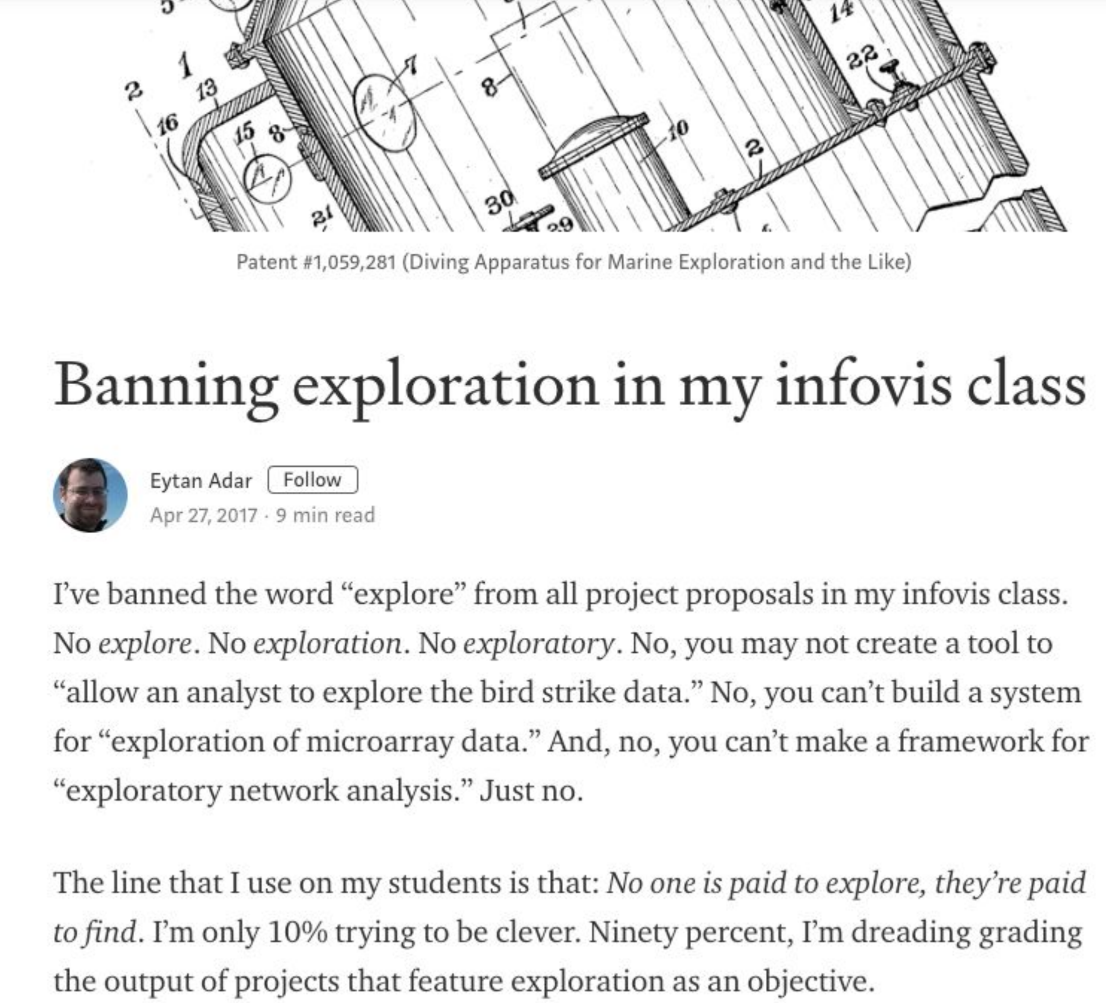


Regularly test your progress by asking a roommate, relative, or friend to force you to flesh out those three steps. Even if you can't take them all confidently, you'll know where you are and where you still have to go. To summarize: Your aim is to explain

1. **Topic:** I am studying __________
2. **Question:** because I want to find out what/why/how __________,
3. **Significance:** in order to help my reader understand

__________.

Example from [Craft of Research, Booth et al]



Patent #1,059,281 (Diving Apparatus for Marine Exploration and the Like)

Banning exploration in my infovis class

Eytan Adar Follow

Apr 27, 2017 · 9 min read

I've banned the word "explore" from all project proposals in my infovis class. No *explore*. No *exploration*. No *exploratory*. No, you may not create a tool to "allow an analyst to explore the bird strike data." No, you can't build a system for "exploration of microarray data." And, no, you can't make a framework for "exploratory network analysis." Just no.

The line that I use on my students is that: *No one is paid to explore, they're paid to find*. I'm only 10% trying to be clever. Ninety percent, I'm dreading grading the output of projects that feature exploration as an objective.

It is important to state the subject of study and expectations of success. Surprise is defined relative to expectations.

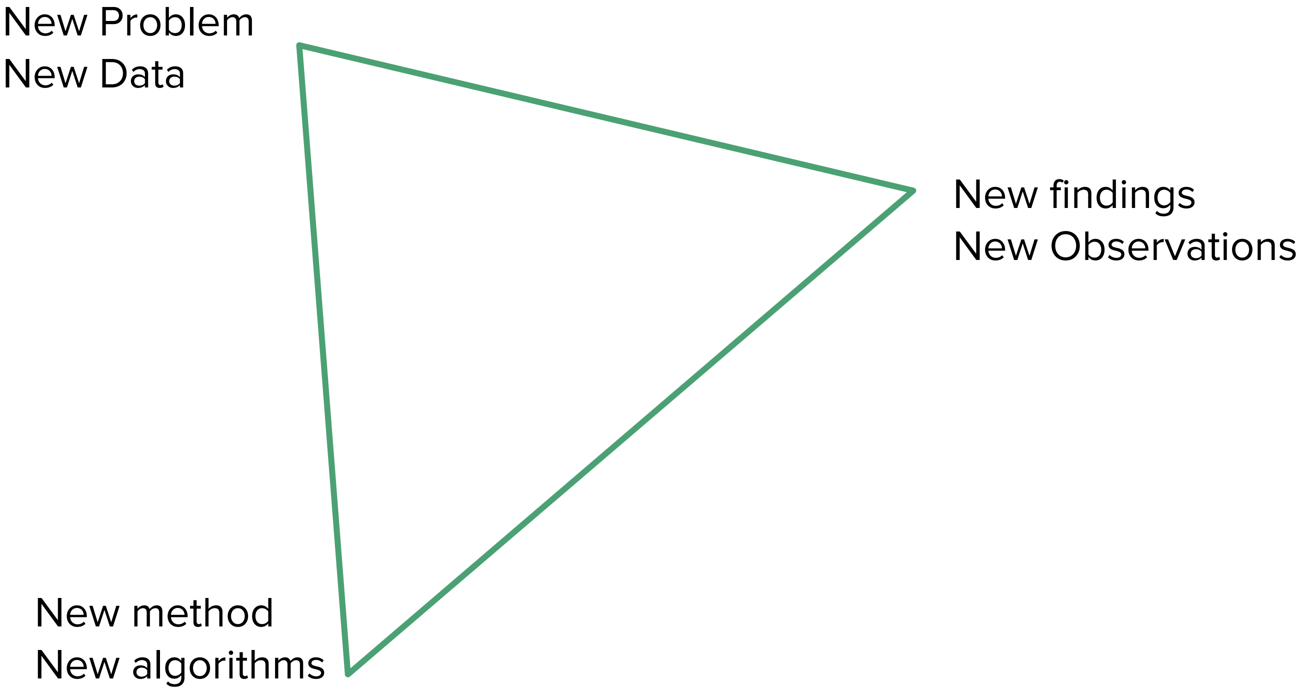
"Framework" is too vague a word to support scientific goals.

https://medium.com/@eytanadar/banning-exploration-in-my-infovis-class-9578676a4705

From questions to tasks

1. Has this (or another highly similar) question been answered?
    a. Google scholar, web search, reports
    b. What is new about my questions / intended approach / intended findings
2. What data
    a. Is my data correct / complete?
    b. What is the nature of the data (if it is new)
    c. Doe this data cover the question you'd like to ask, why or why not
3. What measurements / algorithms
    a. What do people *usually* do for this kind of data
    b. What are the unique things you'd like to do based on your question
4. Are the results correct / expected / surprising?
5. (go back to 1, 2, 3, 4; iterate and repeat)

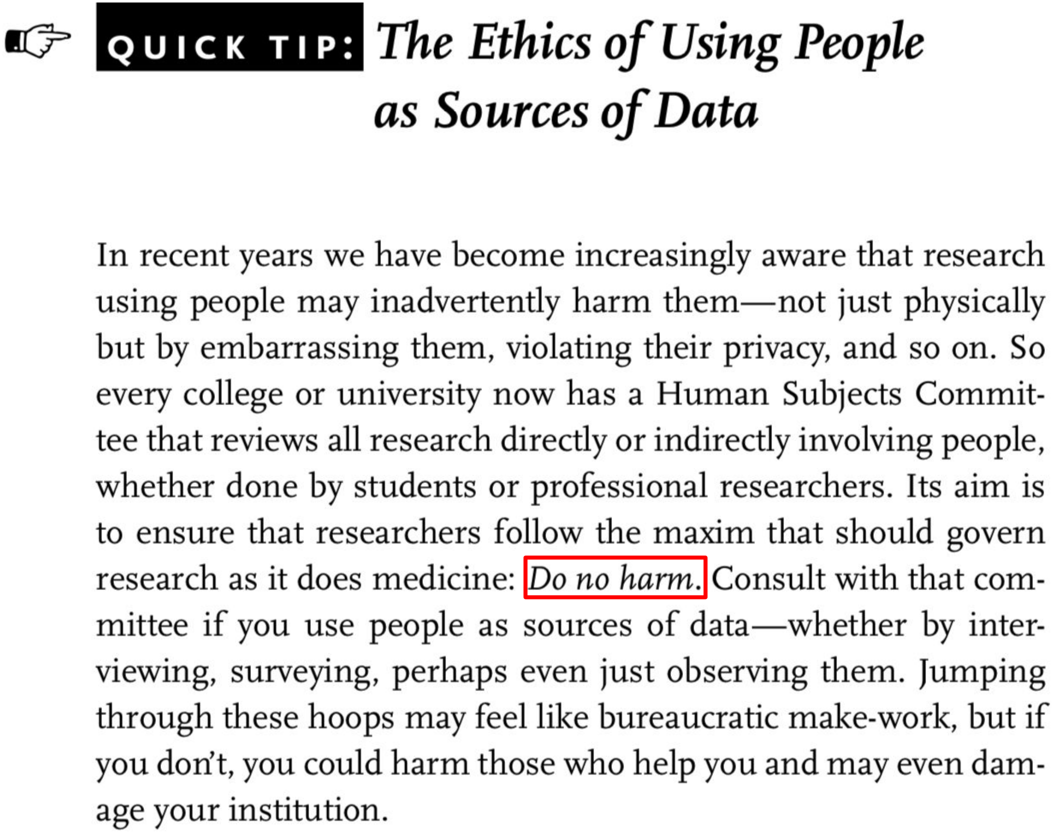
A contribution triangle / polygon



New Problem
New Data

New findings
New Observations

New method
New algorithms

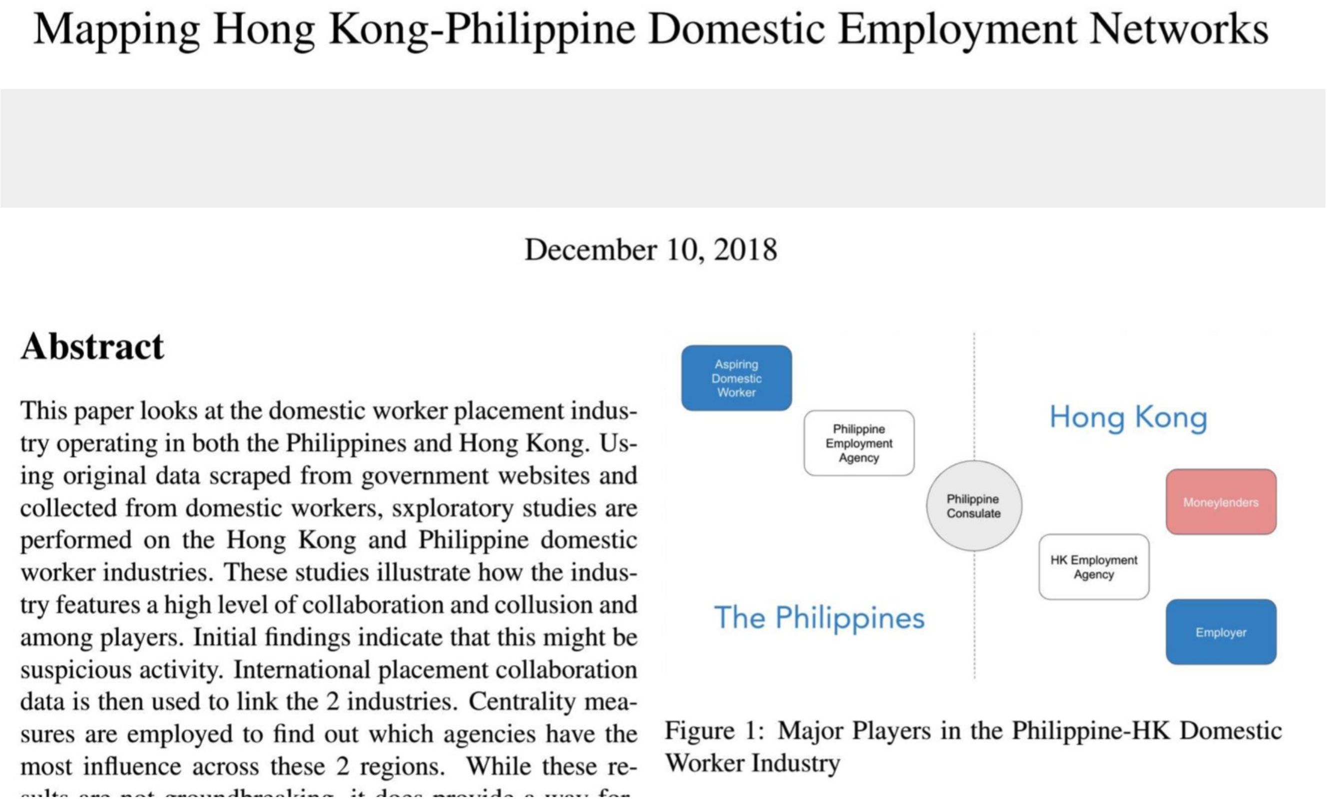


☞ QUICK TIP: *The Ethics of Using People as Sources of Data*

In recent years we have become increasingly aware that research using people may inadvertently harm them—not just physically but by embarrassing them, violating their privacy, and so on. So every college or university now has a Human Subjects Committee that reviews all research directly or indirectly involving people, whether done by students or professional researchers. Its aim is to ensure that researchers follow the maxim that should govern research as it does medicine: *Do no harm.* Consult with that committee if you use people as sources of data—whether by interviewing, surveying, perhaps even just observing them. Jumping through these hoops may feel like bureaucratic make-work, but if you don't, you could harm those who help you and may even damage your institution.

[Craft of Research, Booth et al]

Example 1: HK-Philippenes domestic worker network



Mapping Hong Kong-Philippine Domestic Employment Networks

[redacted]

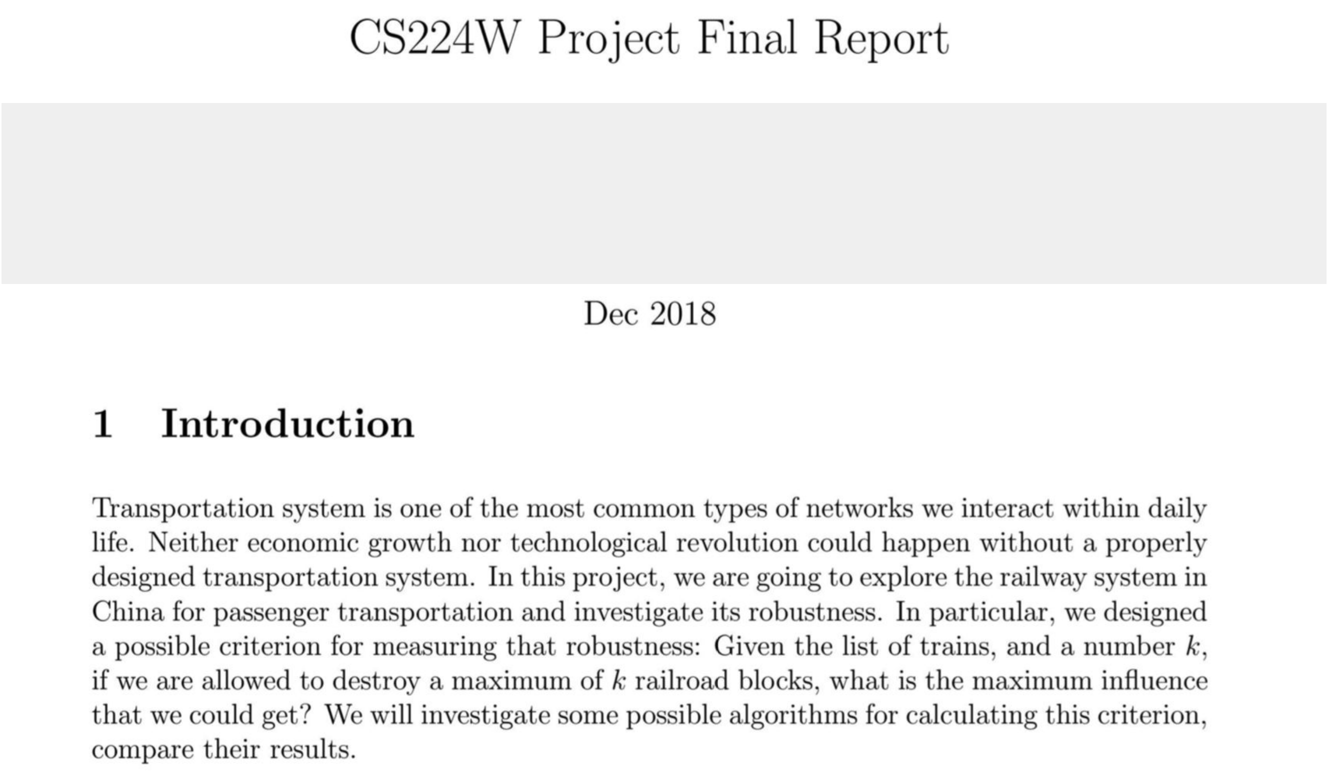
December 10, 2018

Abstract

This paper looks at the domestic worker placement industry operating in both the Philippines and Hong Kong. Using original data scraped from government websites and collected from domestic workers, sxploratory studies are performed on the Hong Kong and Philippine domestic worker industries. These studies illustrate how the industry features a high level of collaboration and collusion and among players. Initial findings indicate that this might be suspicious activity. International placement collaboration data is then used to link the 2 industries. Centrality measures are employed to find out which agencies have the most influence across these 2 regions. While these results are not groundbreaking, it does provide a way for

Figure 1: Major Players in the Philippine-HK Domestic Worker Industry

Project example 2: Robustness of China Rail network



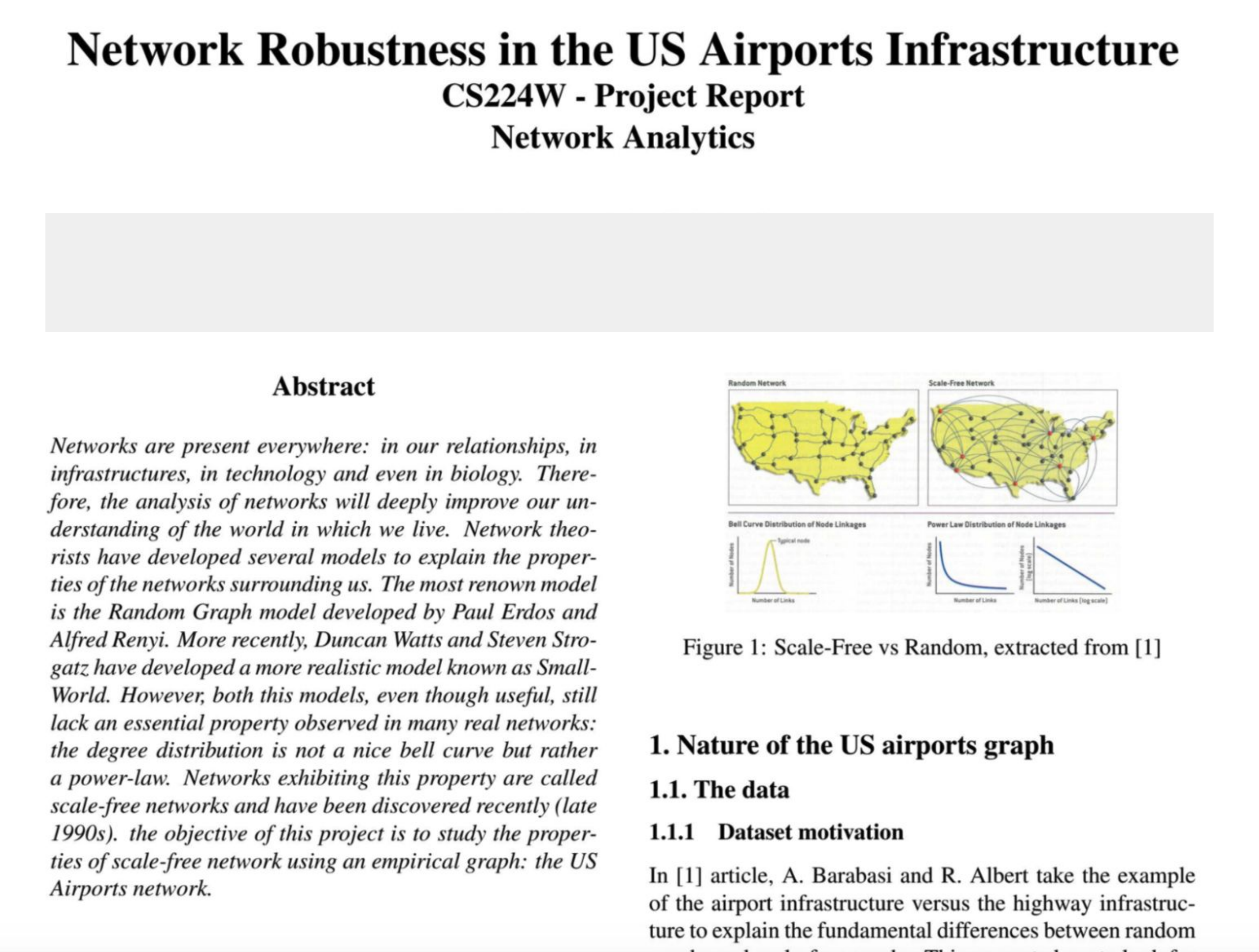
CS224W Project Final Report

[redacted]

Dec 2018

1 Introduction

Transportation system is one of the most common types of networks we interact within daily life. Neither economic growth nor technological revolution could happen without a properly designed transportation system. In this project, we are going to explore the railway system in China for passenger transportation and investigate its robustness. In particular, we designed a possible criterion for measuring that robustness: Given the list of trains, and a number k, if we are allowed to destroy a maximum of k railroad blocks, what is the maximum influence that we could get? We will investigate some possible algorithms for calculating this criterion, compare their results.



Network Robustness in the US Airports Infrastructure

CS224W - Project Report

Network Analytics

Abstract

Networks are present everywhere: in our relationships, in infrastructures, in technology and even in biology. Therefore, the analysis of networks will deeply improve our understanding of the world in which we live. Network theorists have developed several models to explain the properties of the networks surrounding us. The most renown model is the Random Graph model developed by Paul Erdos and Alfred Renyi. More recently, Duncan Watts and Steven Strogatz have developed a more realistic model known as Small-World. However, both this models, even though useful, still lack an essential property observed in many real networks: the degree distribution is not a nice bell curve but rather a power-law. Networks exhibiting this property are called scale-free networks and have been discovered recently (late 1990s). the objective of this project is to study the properties of scale-free network using an empirical graph: the US Airports network.

Random Network
Scale-Free Network
Bell Curve Distribution of Node Linkages
Typical node
Number of Nodes
Number of Links
Power Law Distribution of Node Linkages
Number of Nodes
Number of Links
Number of Nodes (log scale)
Number of Links (log scale)

Figure 1: Scale-Free vs Random, extracted from [1]

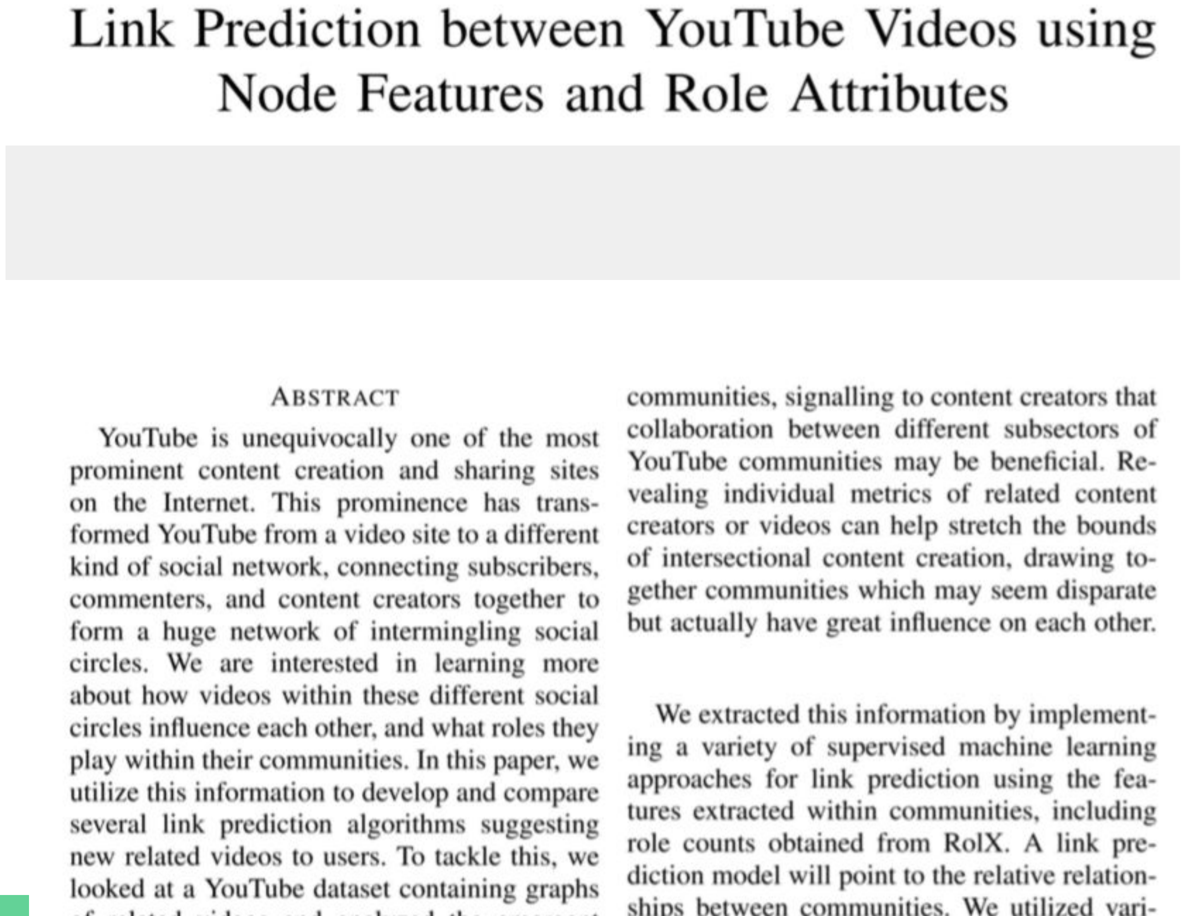
1. Nature of the US airports graph

1.1. The data

1.1.1 Dataset motivation

In [1] article, A. Barabasi and R. Albert take the example of the airport infrastructure versus the highway infrastructure to explain the fundamental differences between random

Project example 3: ML/Recsys on some network data



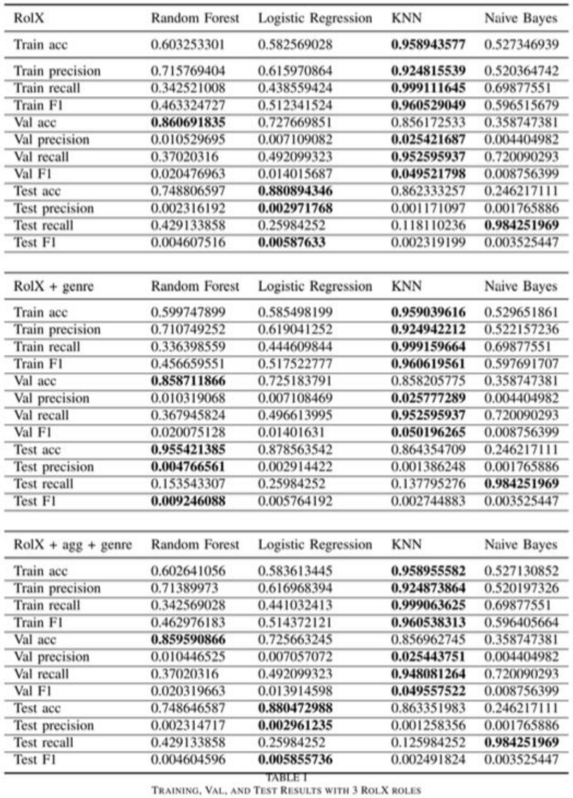
Link Prediction between YouTube Videos using Node Features and Role Attributes

Abstract

YouTube is unequivocally one of the most prominent content creation and sharing sites on the Internet. This prominence has transformed YouTube from a video site to a different kind of social network, connecting subscribers, commenters, and content creators together to form a huge network of intermingling social circles. We are interested in learning more about how videos within these different social circles influence each other, and what roles they play within their communities. In this paper, we utilize this information to develop and compare several link prediction algorithms suggesting new related videos to users. To tackle this, we looked at a YouTube dataset containing graphs

communities, signalling to content creators that collaboration between different subsectors of YouTube communities may be beneficial. Revealing individual metrics of related content creators or videos can help stretch the bounds of intersectional content creation, drawing together communities which may seem disparate but actually have great influence on each other.

We extracted this information by implementing a variety of supervised machine learning approaches for link prediction using the features extracted within communities, including role counts obtained from RolX. A link prediction model will point to the relative relationships between communities. We utilized vari-



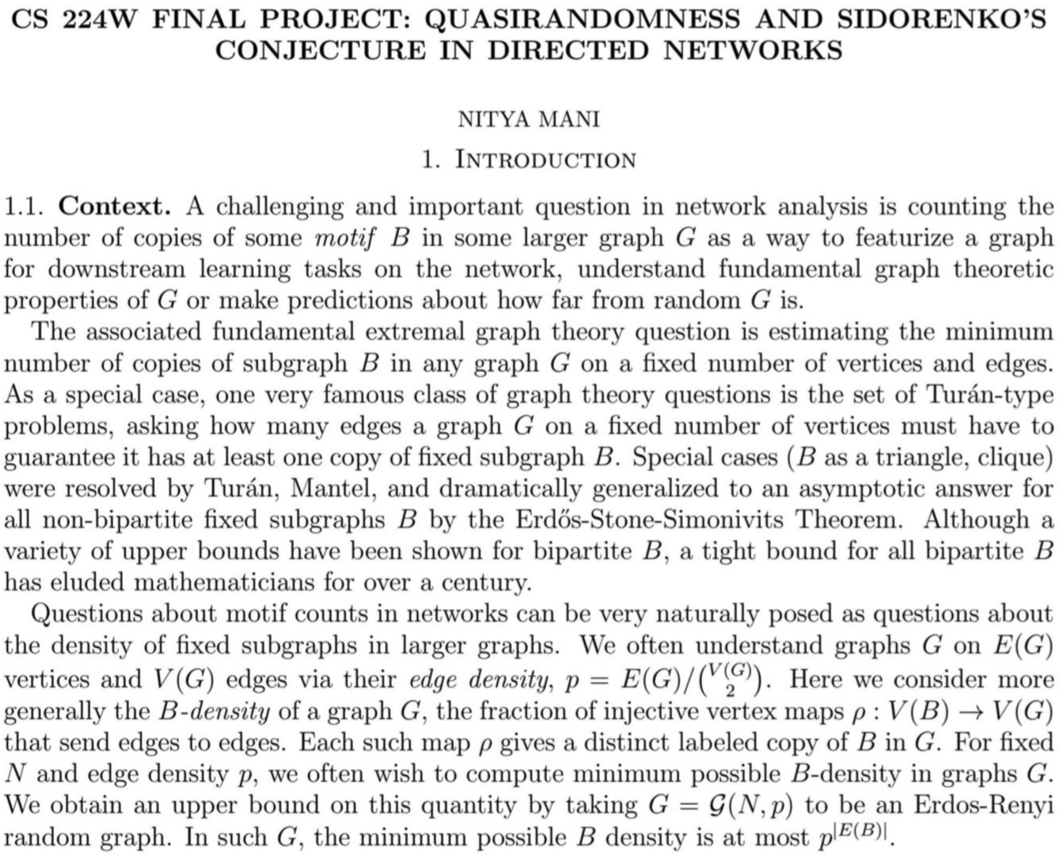
RolX	Random Forest	Logistic Regression	KNN	Naive Bayes
Train acc	0.603253301	0.582569028	**0.958943577**	0.527346939
Train precision	0.715769404	0.615970864	**0.924815539**	0.520364742
Train recall	0.342521008	0.438559424	**0.999111645**	0.69877551
Train F1	0.463324727	0.512341524	**0.960529049**	0.596515679
Val acc	**0.860691835**	0.727669851	0.856172533	0.358747381
Val precision	0.010529695	0.007109082	**0.025421687**	0.004404982
Val recall	0.37020316	0.492099323	**0.952595937**	0.720090293
Val F1	0.020476963	0.014015687	**0.049521798**	0.008756399
Test acc	0.748806597	**0.880894346**	0.862333257	0.246217111
Test precision	0.002316192	**0.002971768**	0.001171097	0.001765886
Test recall	0.429133858	0.25984252	0.118110236	**0.984251969**
Test F1	0.004607516	**0.00587633**	0.002319199	0.003525447

RolX + genre	Random Forest	Logistic Regression	KNN	Naive Bayes
Train acc	0.599747899	0.585498199	**0.959039616**	0.529651861
Train precision	0.710749252	0.619041252	**0.924942212**	0.522157236
Train recall	0.336398559	0.444609844	**0.999159664**	0.69877551
Train F1	0.456659551	0.517522777	**0.960619561**	0.597691707
Val acc	**0.858711866**	0.725183791	0.858205775	0.358747381
Val precision	0.010319068	0.007108469	**0.025777289**	0.004404982
Val recall	0.367945824	0.496613995	**0.952595937**	0.720090293
Val F1	0.020075128	0.01401631	**0.050196265**	0.008756399
Test acc	**0.955421385**	0.878563542	0.864354709	0.246217111
Test precision	**0.004766561**	0.002914422	0.001386248	0.001765886
Test recall	0.153543307	0.25984252	0.137795276	**0.984251969**
Test F1	**0.009246088**	0.005764192	0.002744883	0.003525447

RolX + agg + genre	Random Forest	Logistic Regression	KNN	Naive Bayes
Train acc	0.602641056	0.583613445	**0.958955582**	0.527130852
Train precision	0.71389973	0.616968394	**0.924873864**	0.520197326
Train recall	0.342569028	0.441032413	**0.999063625**	0.69877551
Train F1	0.462976183	0.514372121	**0.960538313**	0.596405664
Val acc	**0.859590866**	0.725663245	0.856962745	0.358747381
Val precision	0.010446525	0.007057072	**0.025443751**	0.004404982
Val recall	0.37020316	0.492099323	**0.948081264**	0.720090293
Val F1	0.020319663	0.013914598	**0.049557522**	0.008756399
Test acc	0.748646587	**0.880472988**	0.863351983	0.246217111
Test precision	0.002314717	**0.002961235**	0.001258356	0.001765886
Test recall	0.429133858	0.25984252	0.125984252	**0.984251969**
Test F1	0.004604596	**0.005855736**	0.002491824	0.003525447

TABLE I
TRAINING, VAL, AND TEST RESULTS WITH 3 ROLX ROLES

Example 4: Quasirandomness



CS 224W FINAL PROJECT: QUASIRANDOMNESS AND SIDORENKO'S CONJECTURE IN DIRECTED NETWORKS

NITYA MANI

1. Introduction

1.1. **Context.** A challenging and important question in network analysis is counting the number of copies of some *motif* B in some larger graph G as a way to featurize a graph for downstream learning tasks on the network, understand fundamental graph theoretic properties of G or make predictions about how far from random G is.

The associated fundamental extremal graph theory question is estimating the minimum number of copies of subgraph B in any graph G on a fixed number of vertices and edges. As a special case, one very famous class of graph theory questions is the set of Turán-type problems, asking how many edges a graph G on a fixed number of vertices must have to guarantee it has at least one copy of fixed subgraph B. Special cases (B as a triangle, clique) were resolved by Turán, Mantel, and dramatically generalized to an asymptotic answer for all non-bipartite fixed subgraphs B by the Erdős-Stone-Simonivits Theorem. Although a variety of upper bounds have been shown for bipartite B, a tight bound for all bipartite B has eluded mathematicians for over a century.

Questions about motif counts in networks can be very naturally posed as questions about the density of fixed subgraphs in larger graphs. We often understand graphs G on $E(G)$ vertices and $V(G)$ edges via their *edge density*, $p = E(G)/\binom{V(G)}{2}$. Here we consider more generally the B-*density* of a graph G, the fraction of injective vertex maps $\rho : V(B) \to V(G)$ that send edges to edges. Each such map ρ gives a distinct labeled copy of B in G. For fixed N and edge density p, we often wish to compute minimum possible B-density in graphs G. We obtain an upper bound on this quantity by taking $G = \mathcal{G}(N, p)$ to be an Erdos-Renyi random graph. In such G, the minimum possible B density is at most $p^{|E(B)|}$.

Project mechanics

- Teaming

  Statement of contributions

- Evaluation criteria

Grading breakdown

- Proposal 5% (week 10)
- Presentation 10% (week 12 Tue, lecture time)
- Report 35% (end of week 12)

Solo or 2-person teams

Students are encouraged to work alone or in a team of 2.
Grades will take into account the number of people involved and the results achieved.

Statement of contributions:
The project final report is expected to include a one-sentence statement of who-did-what.

What we look for in grading the final report:

The technical quality of the work: Does the technical material make sense? Are the methods tried reasonable? Are the proposed algorithms or applications clever and interesting? Do the authors convey novel insights about the problem and/or algorithms?

Significance: Did the authors choose an interesting or a "real" problem to work on, or only a small "toy" problem? Is this work likely to be useful and/or have impact?

The novelty of the work.

The clarity of the write-up.

Project proposal - 1 page
due week 10, start thinking about it now

- What is the problem you are solving?
- What data will you use (how will you get it)?

- Which algorithms/techniques/models you plan to use/develop? Be as specific as you can!
- Who will you evaluate your method? How will you test it? How will you measure success?
- What results do you expect?