Ramsey Theory and Big Data

Micheal Pawliuk

University of Calgary

November 19, 2018

Joint Work with M.Waddell (Columbia University)

Stories

1. Story 1: Motivations
2. Story 2: Goodman's Theorem
3. Story 3: How many bad colourings?

Beginning of the story

Calude and Longo, 2016: "The Deluge of Spurious Correlations in Big Data"

Hey, data scientists and statisticians, Ramsey Theory should say something about large data sets.

Beginning of the story

Calude and Longo, 2016: "The Deluge of Spurious Correlations in Big Data"

Hey, data scientists and statisticians, Ramsey Theory should say something about large data sets.

M. Waddell (Columbia, PhD student in data science)

Hey, Mike, let's follow the lead of Calude-Longo.

Why develop these connections?

- The connections are natural.

Why develop these connections?

- The connections are natural.
- We know Ramsey technology the best.

Why develop these connections?

- The connections are natural.
- We know Ramsey technology the best.
- The connections are largely unexplored.

Why develop these connections?

- The connections are natural.
- We know Ramsey technology the best.
- The connections are largely unexplored.
- Impactful, applicable research.

Story 2: Goodman's Theorem

Spurious Correlations

Toy Problem 1

You discover that on Tuesday, Honza wore 3 shirts. (You also know that he wore 11 shirts over the course of the 5-day conference.)

Should we conclude that something special happened to Honza on Tuesday?

Spurious Correlations

Toy Problem 1

You discover that on Tuesday, Honza wore 3 shirts. (You also know that he wore 11 shirts over the course of the 5-day conference.)
Should we conclude that something special happened to Honza on Tuesday?

Toy Problem 2

You discover that for a particular red-blue edge colouring of K_{100}, about 30% of the triangles are monochromatic.
How important is this?

Spurious Correlations

Toy Problem 1

You discover that on Tuesday, Honza wore 3 shirts. (You also know that he wore 11 shirts over the course of the 5-day conference.)
Should we conclude that something special happened to Honza on Tuesday?

Toy Problem 2

You discover that for a particular red-blue edge colouring of K_{100}, about 30% of the triangles are monochromatic.
How important is this?

A **spurious correlation** is one that is a result of forced, geometric or combinatorial relations.

$K(3,3) = 6$

Theorem (Ramsey 1929)

Fix a ("small") m. There is a ("large") n such that *every* blue/red edge colouring of K_n contains a monochromatic K_m.

$K(3,3) = 6$

Theorem (Ramsey 1929)

Fix a ("small") m. There is a ("large") n such that *every* blue/red edge colouring of K_n contains a monochromatic K_m.

Putnam Contest 1953

Every red/blue edge colouring of a K_6 contains a monochromatic triangle.

$K(3,3) = 6$

Theorem (Ramsey 1929)

Fix a ("small") m. There is a ("large") n such that *every* blue/red edge colouring of K_n contains a monochromatic K_m.

Putnam Contest 1953

Every red/blue edge colouring of a K_6 contains a monochromatic triangle.

Theorem (folklore?)

Every red/blue edge colouring of a K_6 contains <u>at least 2</u> monochromatic triangle.

Enter Goodman

Theorem (Goodman 1959)

"At least a quarter of the triangles must be monochromatic."

Enter Goodman

Theorem (Goodman 1959)

"At least a quarter of the triangles must be monochromatic."

Theorem (Goodman 1959)

Every red/blue edge colouring of K_n must contain at least $\frac{1}{4}\frac{n-4}{n-1}$ fraction of monochromatic triangles.

Recall: $\binom{5}{3} = 10$, $\binom{6}{3} = 20$.

Enter Goodman

Theorem (Goodman 1959)

"At least a quarter of the triangles must be monochromatic."

Theorem (Goodman 1959)

Every red/blue edge colouring of K_n must contain at least $\frac{1}{4}\frac{n-4}{n-1}$ fraction of monochromatic triangles.

Recall: $\binom{5}{3} = 10$, $\binom{6}{3} = 20$.

Data Analysis

Plan

Take a large red/blue edge coloured graph and measure the percentage of monochromatic triangles.

Data Analysis

Plan

Take a large red/blue edge coloured graph and measure the percentage of monochromatic triangles.

Question 1

Is the observed percentage near 25% or significantly above?

Data Analysis

Plan

Take a large red/blue edge coloured graph and measure the percentage of monochromatic triangles.

Question 1

Is the observed percentage near 25% or significantly above?

Question 2

What is this measuring?

What about K_4?

Conjecture: Erdős

For large n, every red/blue edge coloured K_n, (asymptotically) at least $\frac{1}{32}$ many of the K_4 should be monochromatic.

What about K_4?

Conjecture: Erdős

For large n, every red/blue edge coloured K_n, (asymptotically) at least $\frac{1}{32}$ many of the K_4 should be monochromatic.

Conjecture: Erdős

For large n, every red/blue edge coloured K_n, (asymptotically) at least $\frac{1}{32}$ many of the K_4 should be monochromatic.

What about K_4?

Conjecture: Erdős

For large n, every red/blue edge coloured K_n, (asymptotically) at least $\frac{1}{32}$ many of the K_4 should be monochromatic.

Conjecture: Erdős

For large n, every red/blue edge coloured K_n, (asymptotically) at least $\frac{1}{32}$ many of the K_4 should be monochromatic.

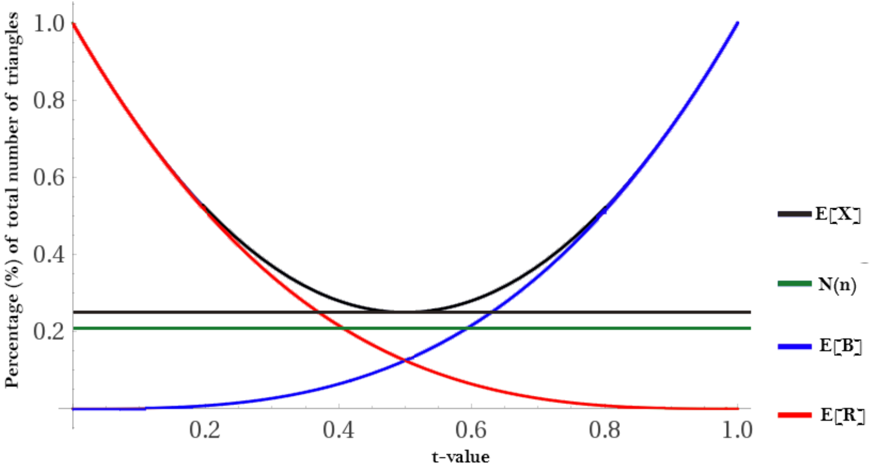
Thomason, 1989

There are red/blue edge colourings of (large) K_n with only $\frac{1}{33}$ many monochromatic triangles.

For K_m: there are colourings with $0.936 \cdot 2^{1-\binom{m}{2}}$ monochromatic K_m.

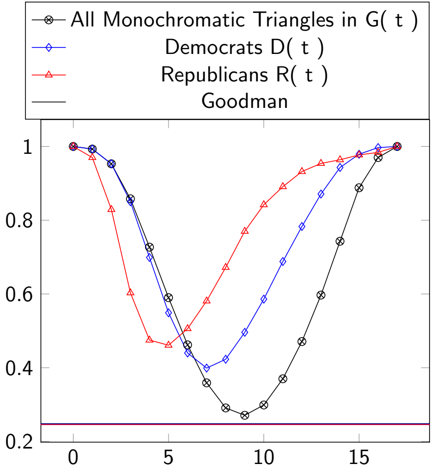
How can this be used?

This can be used to give a meaningful measure of randomness.



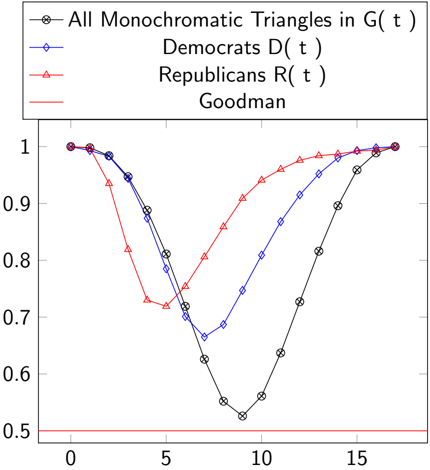
Real data set: 1984 US Congress voting

168 Republicans + 267 Democrats = 435 Voters. 16 votes, Hamming distance.



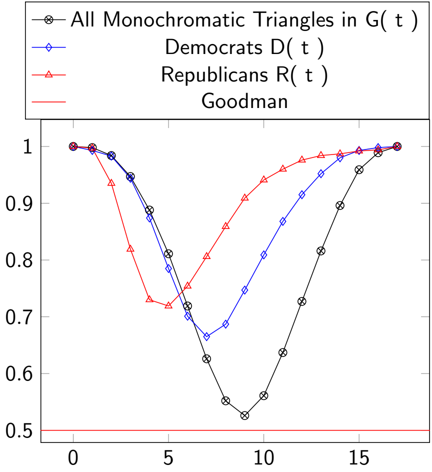
Advantages

1. It is computationally efficient to count triangles (but not K_4s).



Advantages

1. It is computationally efficient to count triangles (but not K_4s).
2. This picture is also a measure of how transitive the combined relations are.



Other "Goodman" theorems

1. Schur triples (2 colours). $\frac{1}{22}$. Story starts with Graham-Rödl-Ruciński 1996, "ends" with Robertson-Zeilberger 2003.
2. VdW (3 term, 2 colours). At least 25% of all 3-term such arithmetic progressions must be monochromatic. [Sjöland 2014, using Cameron-Cilleruelo-Serra 2007]
3. VdW (4 term, 2 colours). $\frac{7}{96} < \frac{1}{16}$. [Lu-Peng 2012, building off Wolf 2010]
4. See also work of Parillo-Robertson-Saracin 2008, Butler-Costello-Graham 2010.

Questions

Question 1

What are the other Goodman (quantitative) versions of Ramsey theorems?

Questions

Question 1

What are the other Goodman (quantitative) versions of Ramsey theorems?

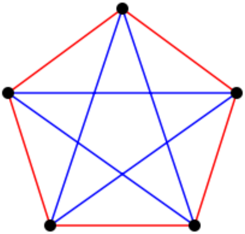
Question 2

What is a "physical" interpretation of monochromatic arithmetic progressions in a large data set?

Story 3: Bad colourings

Artist: M. Pawliuk (Age: 31).

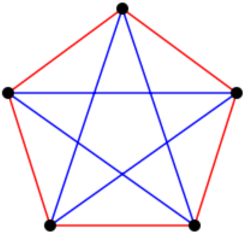
$R(3,3) > 5$



Observation

There is an edge colouring of K_5 without a monochromatic K_3, but *most* edge colourings *do* have a monochromatic triangle.

$R(3,3) > 5$



Observation

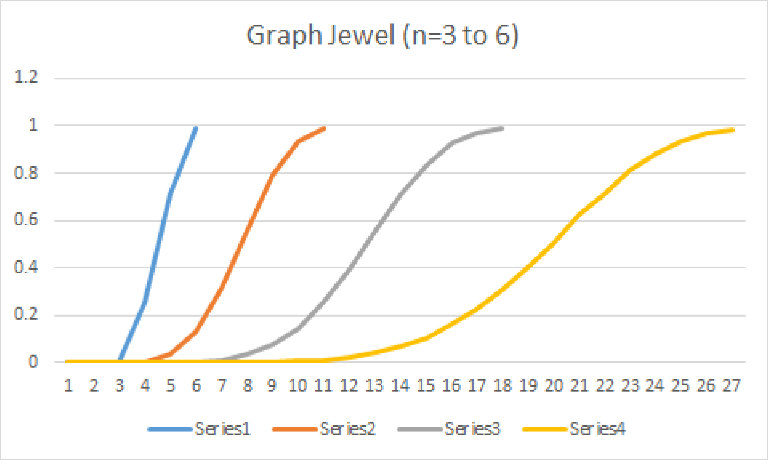
There is an edge colouring of K_5 without a monochromatic K_3, but *most* edge colourings *do* have a monochromatic triangle.

[

noframenumbering]Major Question

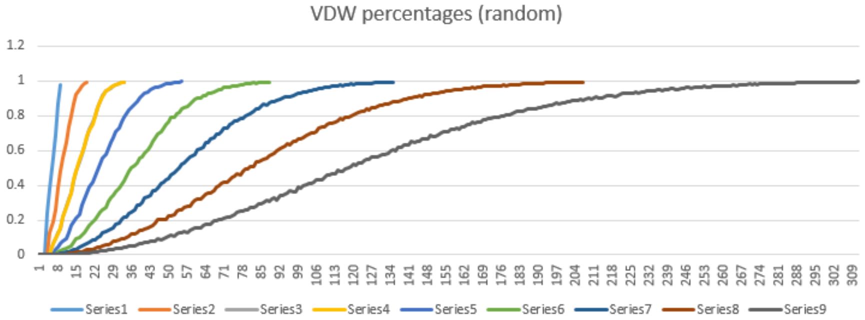
Given a Ramsey-style result, as the size n of the data set grows, what percentage of colourings have monochromatic witnesses?

Ramsey's Theorem



See Robertson-Cipolli-Dascălu 2017 for descriptions of these distributions.

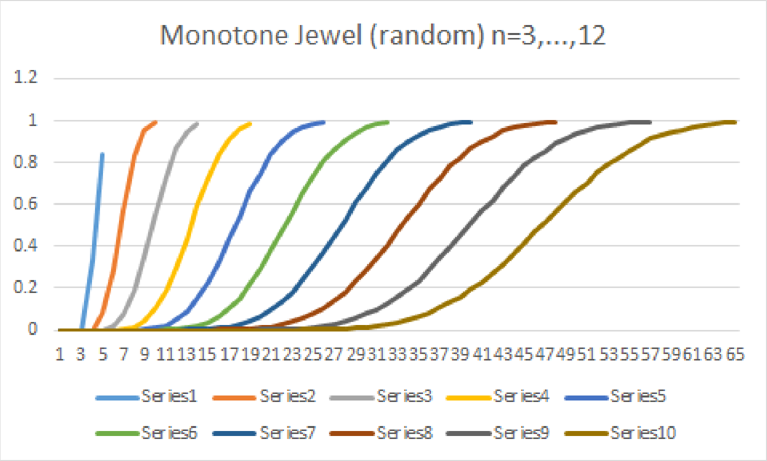
VdW. Arithmetic progressions of length n



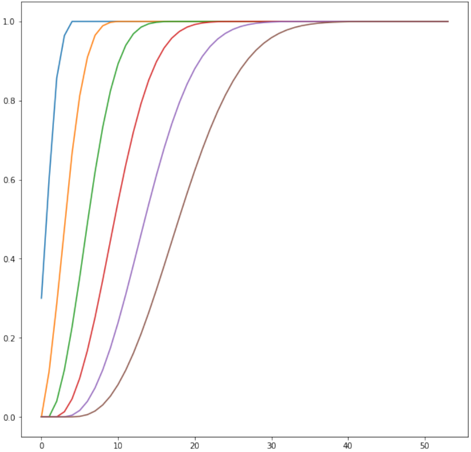
AP of lengths 3 to 10.

See Robertson-Cipolli-Dascălu 2017 for descriptions of these distributions.

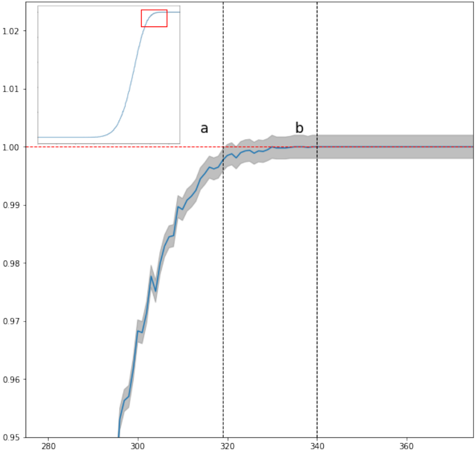
Erdős-Szekeres. Monotone subsequence of length n



Partitions of n objects into N boxes, with at least one box with N objects

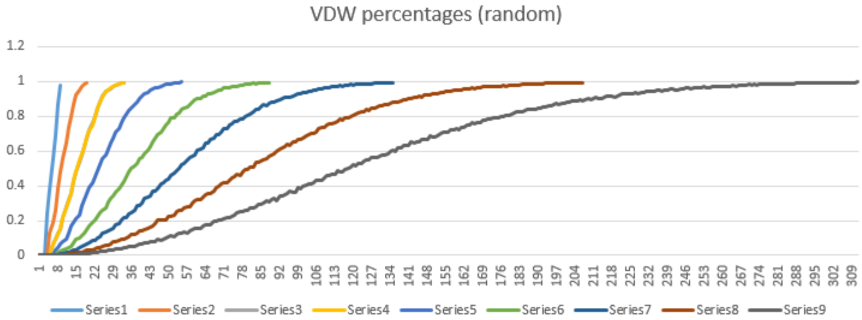


$N = 20$



The jaggedness is not noise! It is an essential feature of the graph.

Machine learning and classifiers



Using **one** of these distributions gives you an okay way to classify/partition graphs. Using **many** of these distributions gives you a better way to classify graphs.

Questions

Question 1

Can these distributions be written in the form "Nice" + "Small"? Where "small" is because of something essential to the geometry of the structures?

Questions

Question 1

Can these distributions be written in the form "Nice" + "Small"? Where "small" is because of something essential to the geometry of the structures?

Question 2

What are the "99% Ramsey numbers" for various Ramsey structures?

Questions

Question 1

Can these distributions be written in the form "Nice" + "Small"? Where "small" is because of something essential to the geometry of the structures?

Question 2

What are the "99% Ramsey numbers" for various Ramsey structures?

Call to action 1

Results in this area need to be made accessible to data scientists. We need to (if possible) include digestible results.

Questions

Question 1

Can these distributions be written in the form "Nice" + "Small"? Where "small" is because of something essential to the geometry of the structures?

Question 2

What are the "99% Ramsey numbers" for various Ramsey structures?

Call to action 1

Results in this area need to be made accessible to data scientists. We need to (if possible) include digestible results.

Call to action 2

Talk to a statistician and a data scientist.