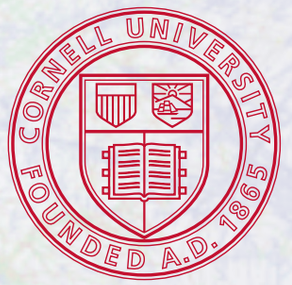
Subgraph Frequencies:
The Empirical and Extremal Geography of Large Graph Collections

Johan Ugander, Lars Backstrom, Jon Kleinberg
NetSci, June 3, 2014
Paper: Proceedings of WWW'13



Cornell University

Graph Collections

- **Neighborhoods:** graph induced by friends of a single ego, excluding ego
- **Groups:** graph induced by members of a Facebook ‘group‘
- **Events:** graph induced by ‘Yes’ respondents to a Facebook ‘event’

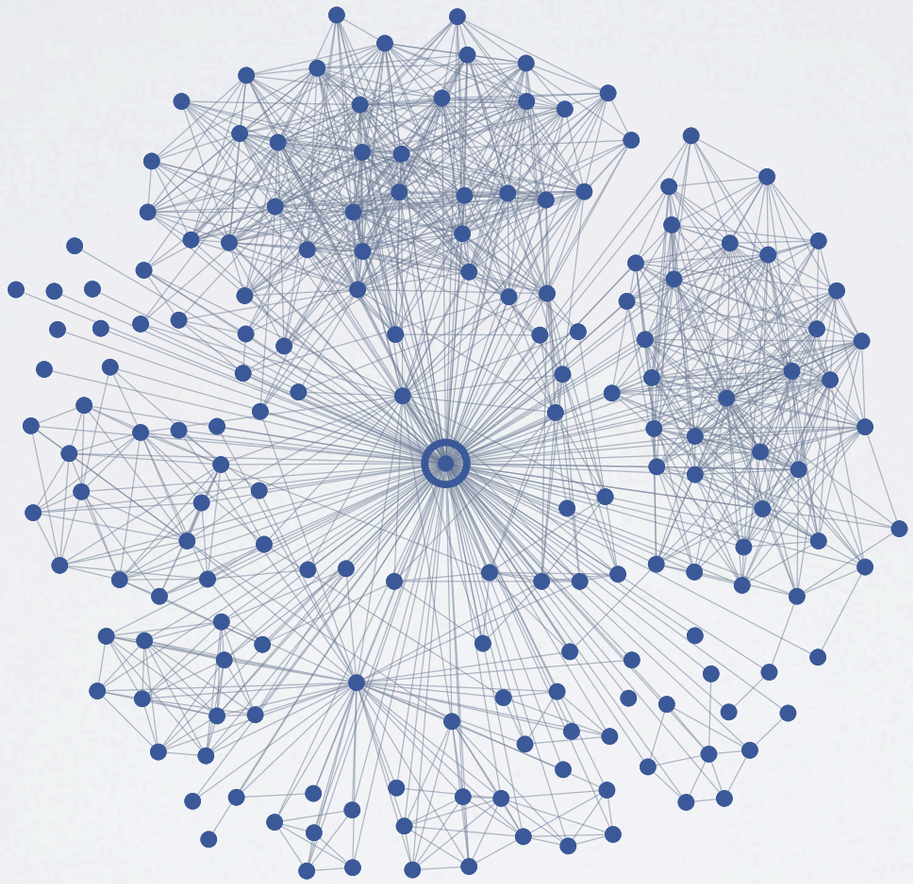


Figure: L. Byron, T. Lento, I. Rosenn, C. Marlow, ‘Maintained Relationships on Facebook’ (overstated.net)

Collections: Faust 2007, Faust 2010, Goel et al 2013, Cheng et al. 2014

Graph Collections

- **Neighborhoods:** graph induced by friends of a single ego, excluding ego
- **Groups:** graph induced by members of a Facebook ‘group‘
- **Events:** graph induced by ‘Yes’ respondents to a Facebook ‘event’

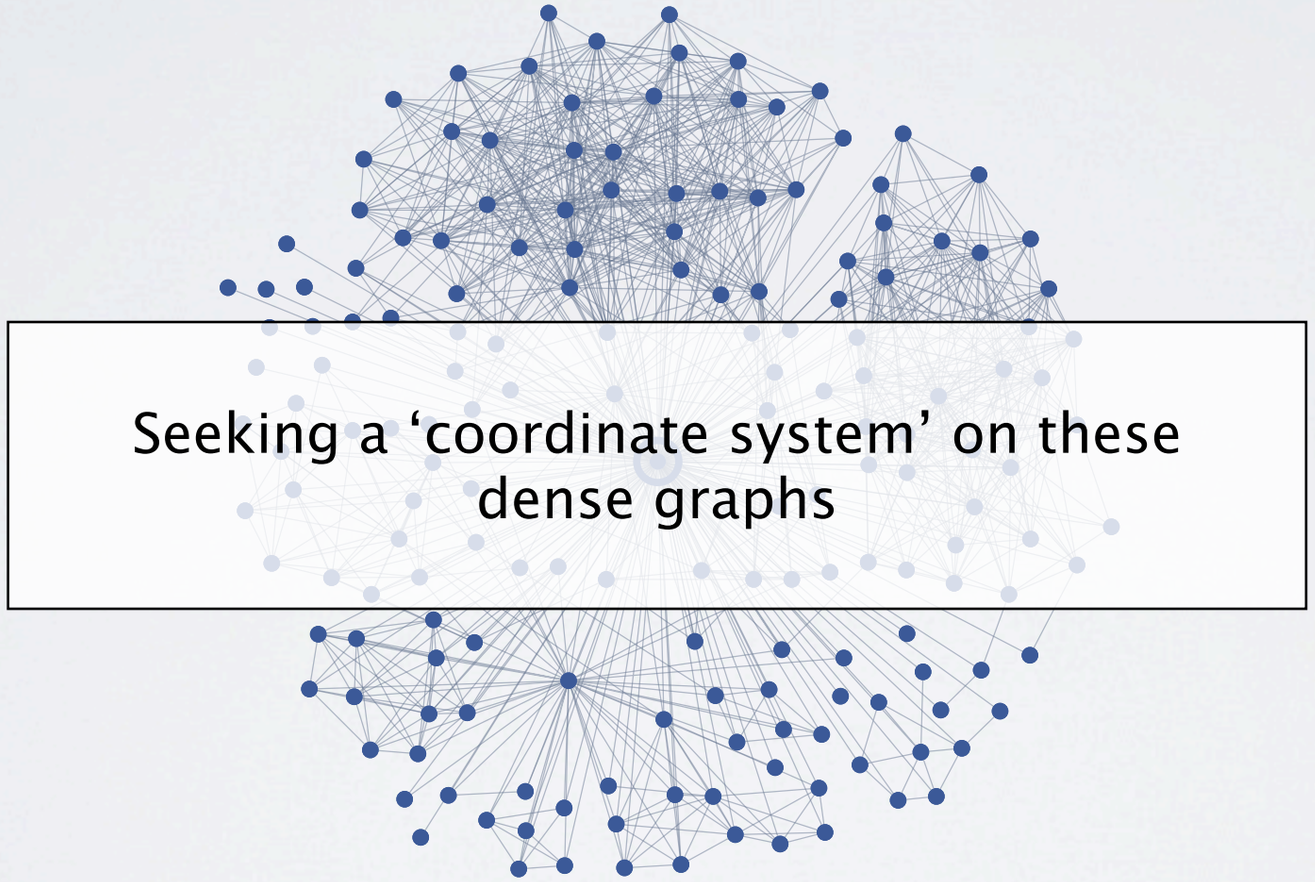
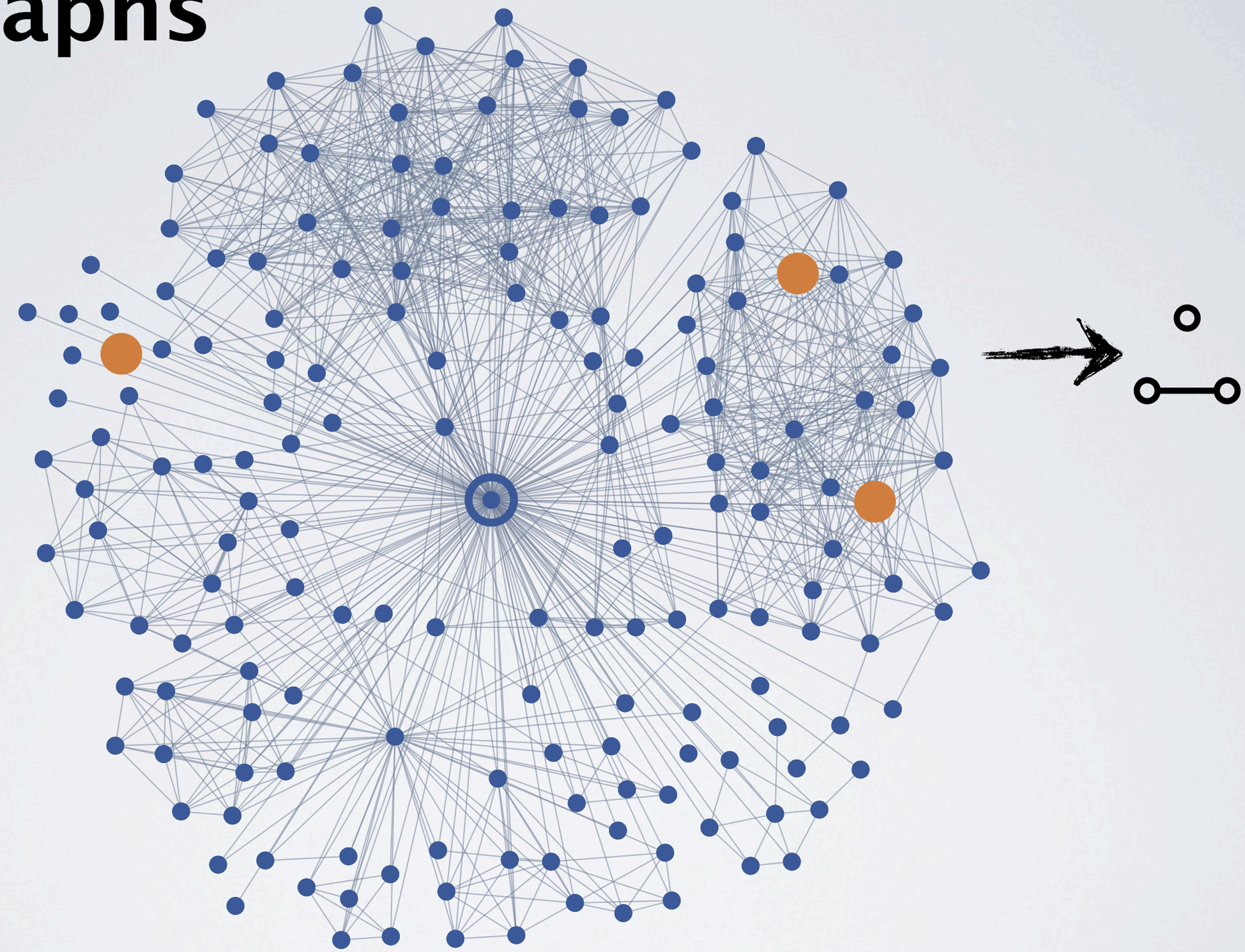


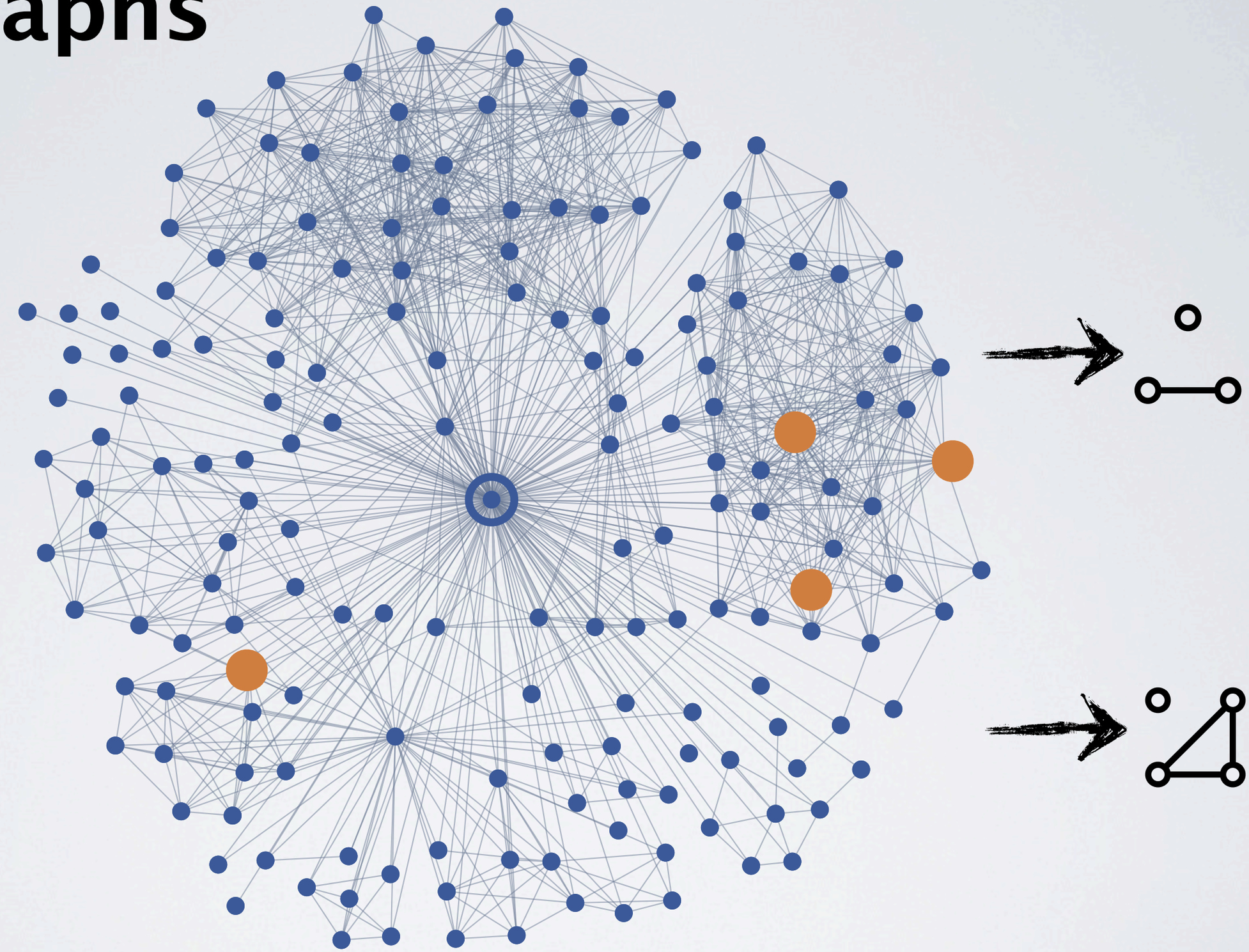
Figure: L. Byron, T. Lento, I. Rosenn, C. Marlow, ‘Maintained Relationships on Facebook’ (overstated.net)

Collections: Faust 2007, Faust 2010, Goel et al 2013, Cheng et al. 2014

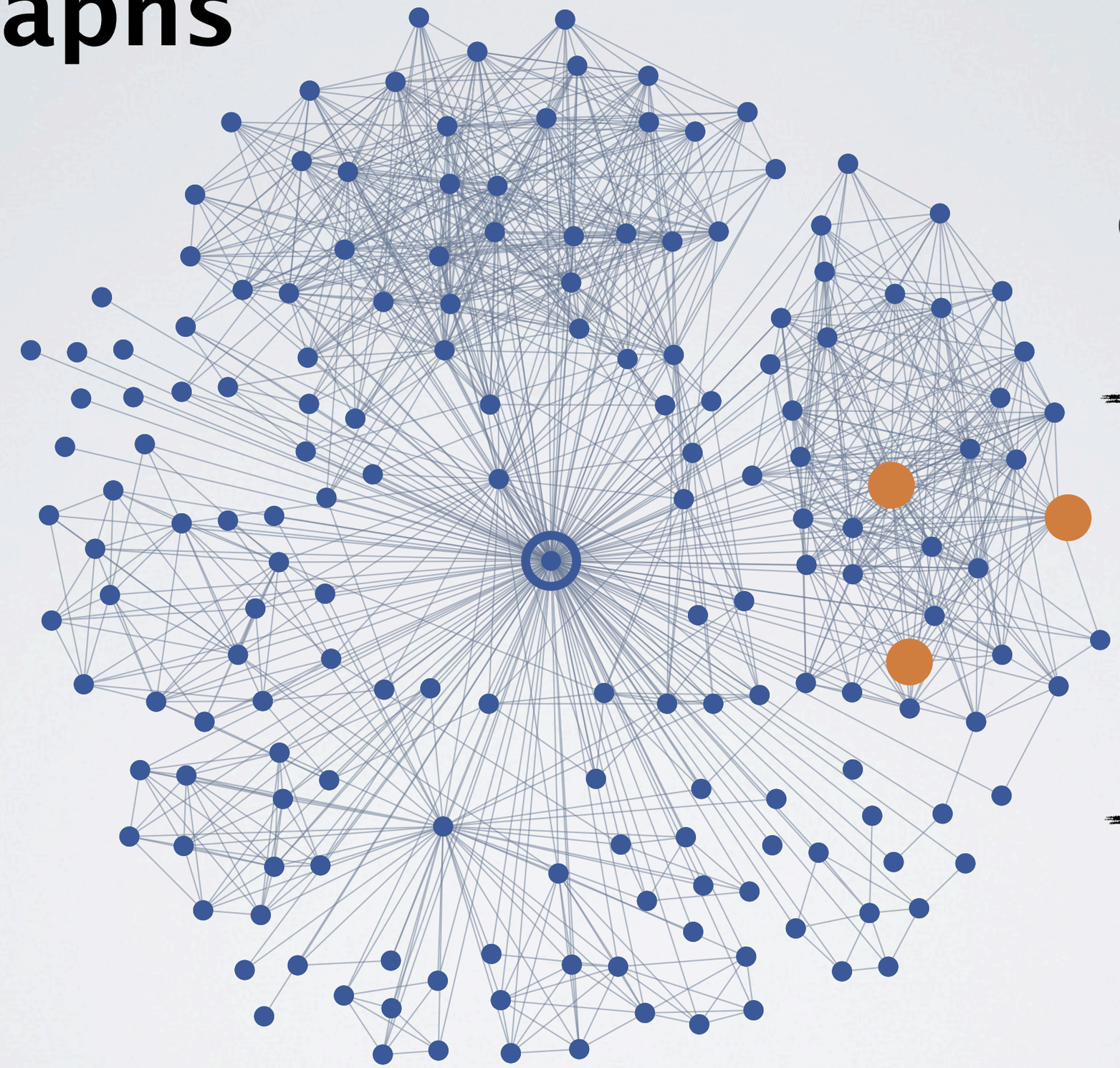
Subgraphs



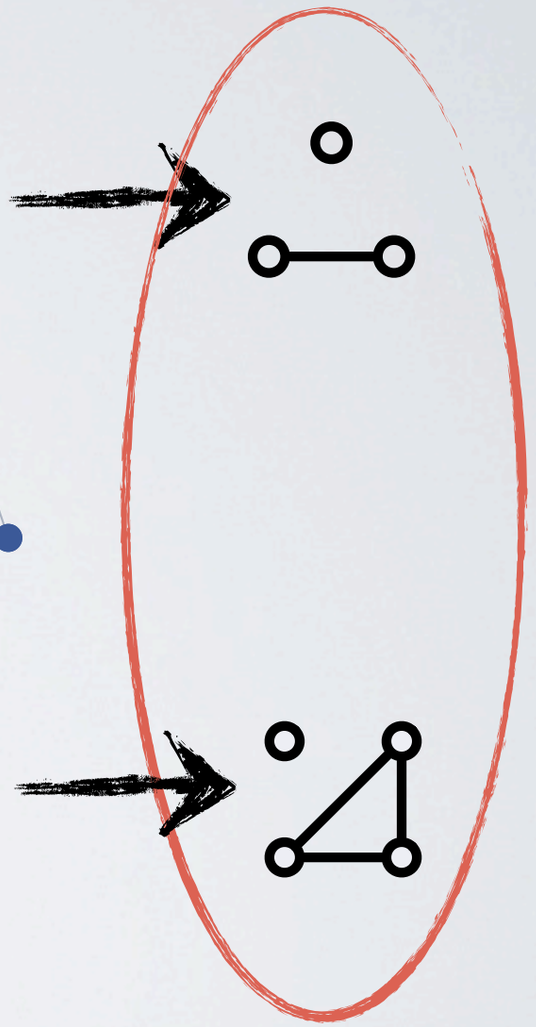
Subgraphs



Subgraphs



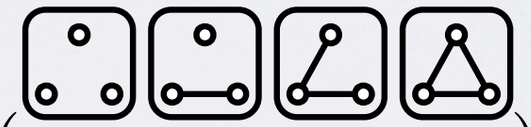
Compute frequencies

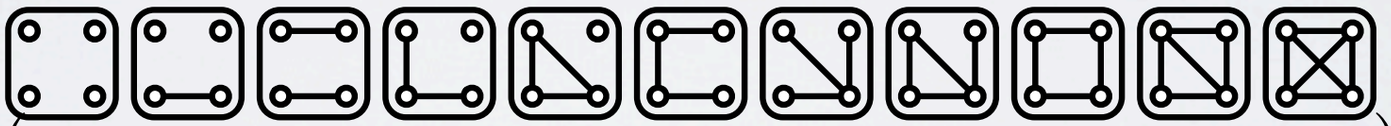


Subgraph Frequencies

- **Definition:** The **subgraph frequency s(F,G)** of a k–node subgraph **F** in a graph **G** is the fraction of k–tuples of nodes in **G** that induce a copy of **F.**

- **Subgraph frequency vectors:**

$s(\cdot, G) = ($  $) = (x_1,\ x_2,\ x_3,\ x_4)$ $=$ (0.18, 0.37, 0.14, 0.31)

$s(\cdot, G) = ($  $) = (y_1,\ y_2,\ y_3,\ y_4,\ y_5,\ y_6,\ y_7,\ y_8,\ y_9,\ y_{10},\ y_{11})$

Triad census: Davis–Leinhardt 1971, Wasserman–Faust 1994

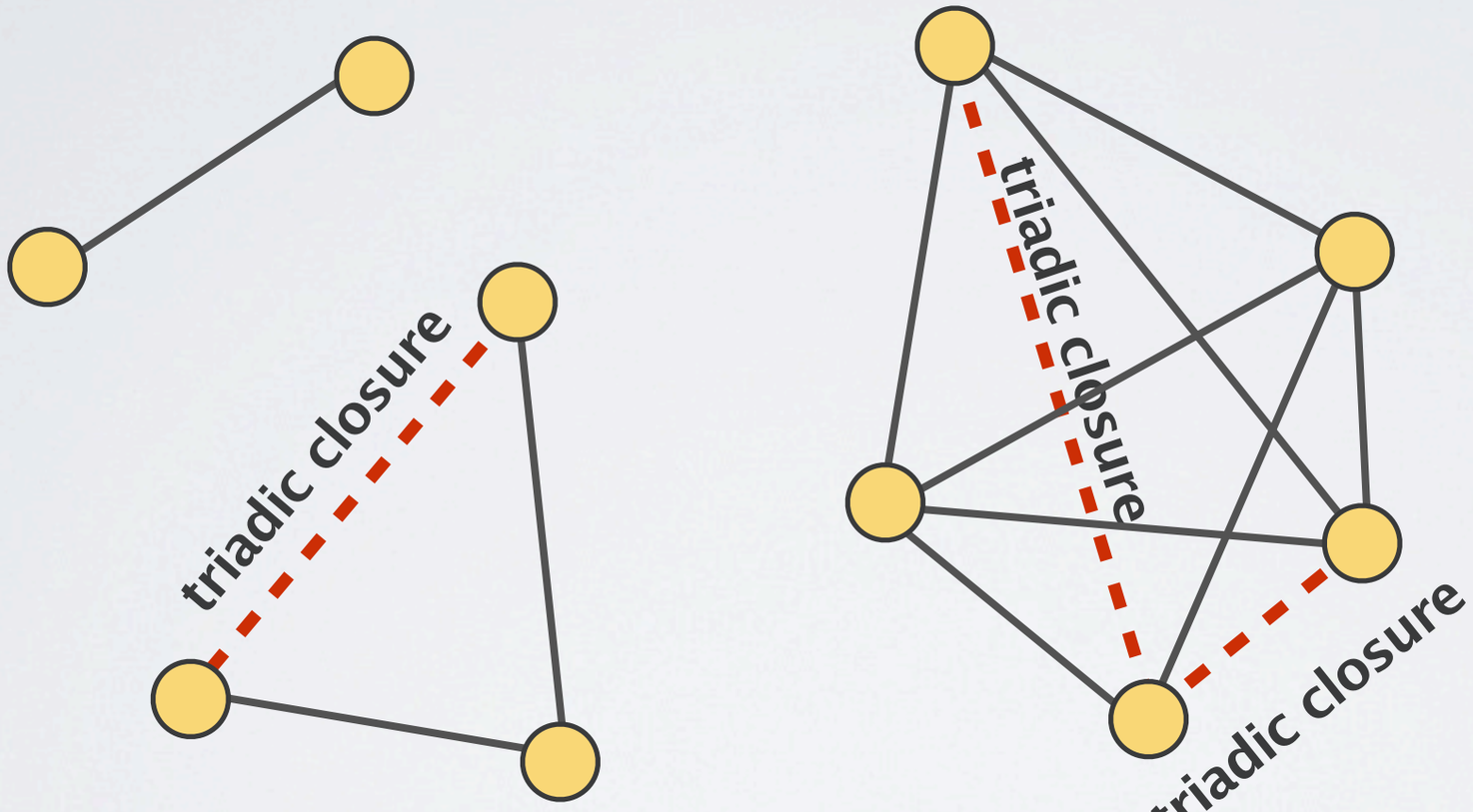
Motifs/Frequent subgraphs: Inokuchi et al. 2000, Milo et al. 2002, Kuramochi–Karypis 2004, Adamic et al. 2008

Empirical vs. Extremal

- Consider the subgraph frequencies space:
  - **Empirical Geography:** **Where do social graphs live?**
  - **Extremal Geography:** **What's feasible, combinatorially?**

What's a property of people and what's a property of graphs?

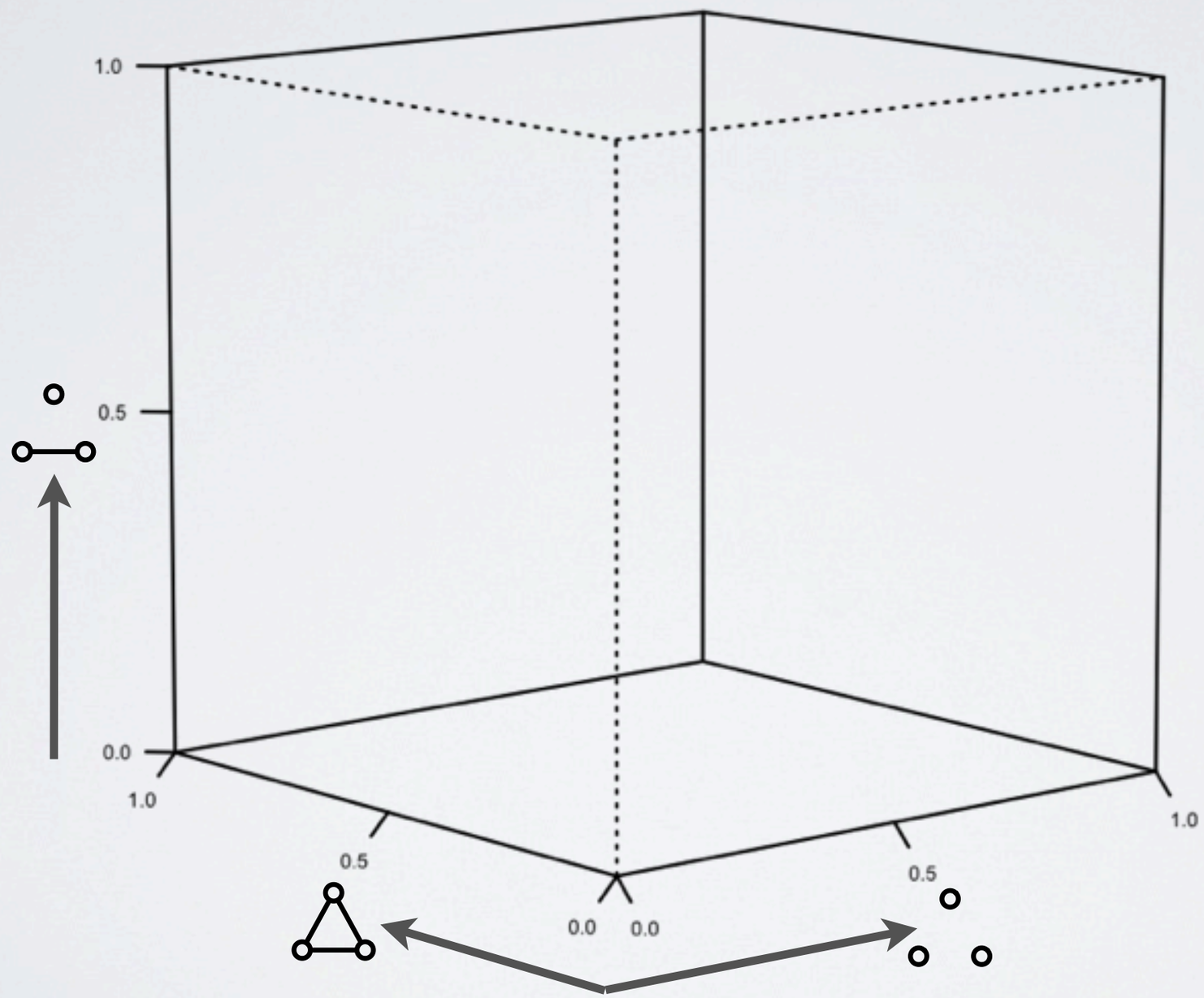
What do we expect?



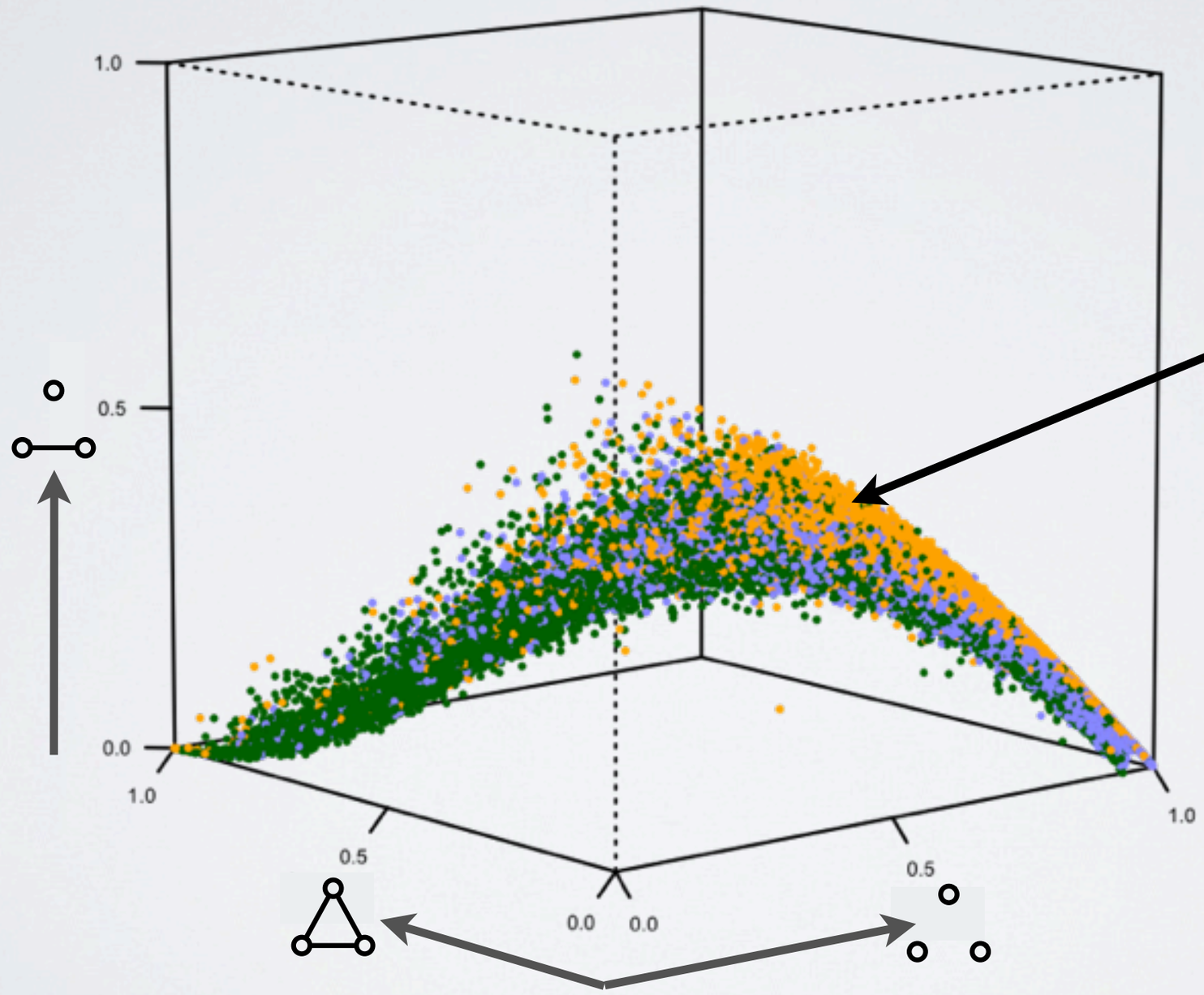
Expect few wedges, many triangles.

Triadic closure: Rapoport 1953

The triad space

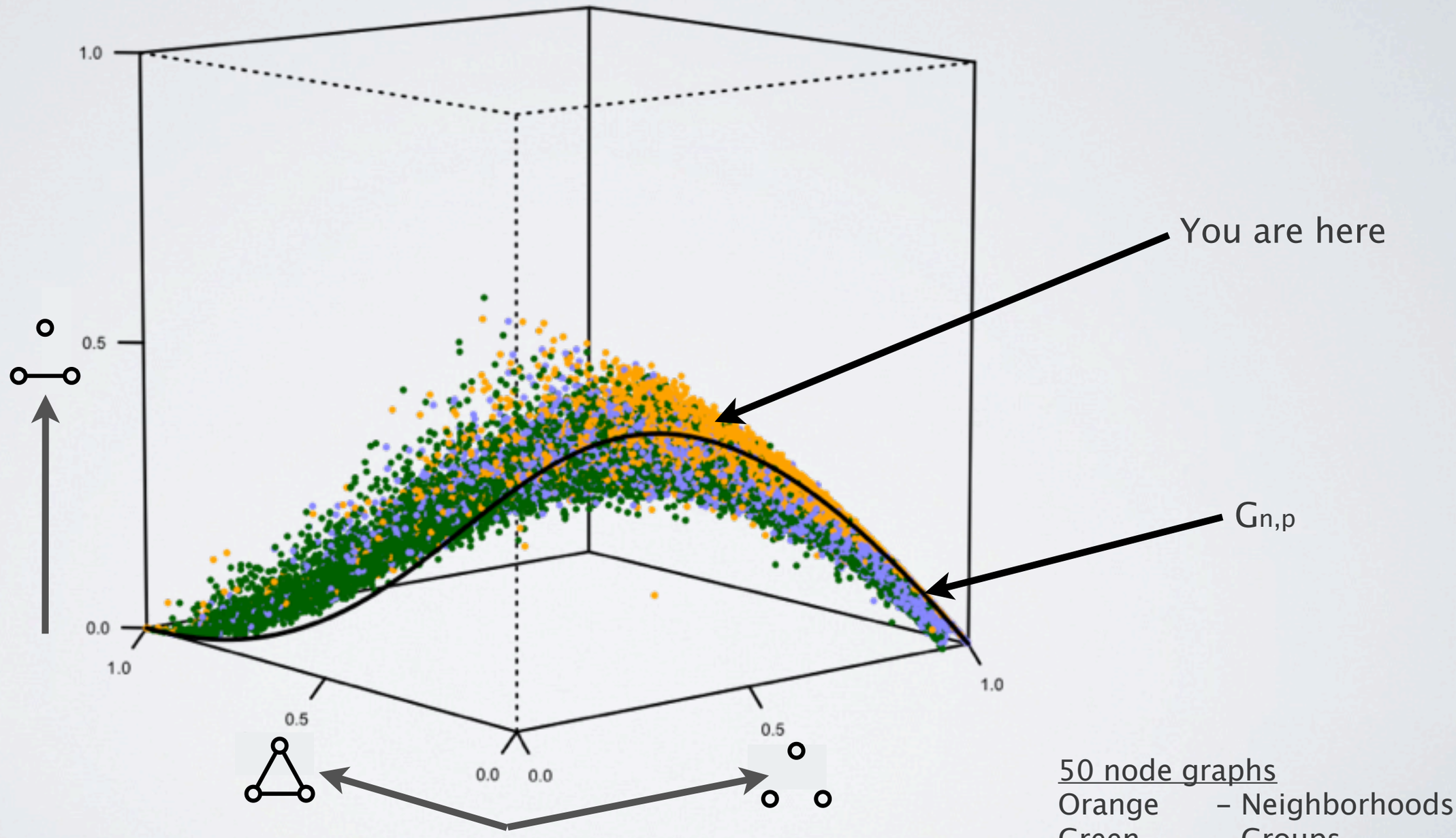


The triad space

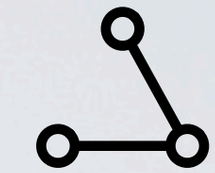


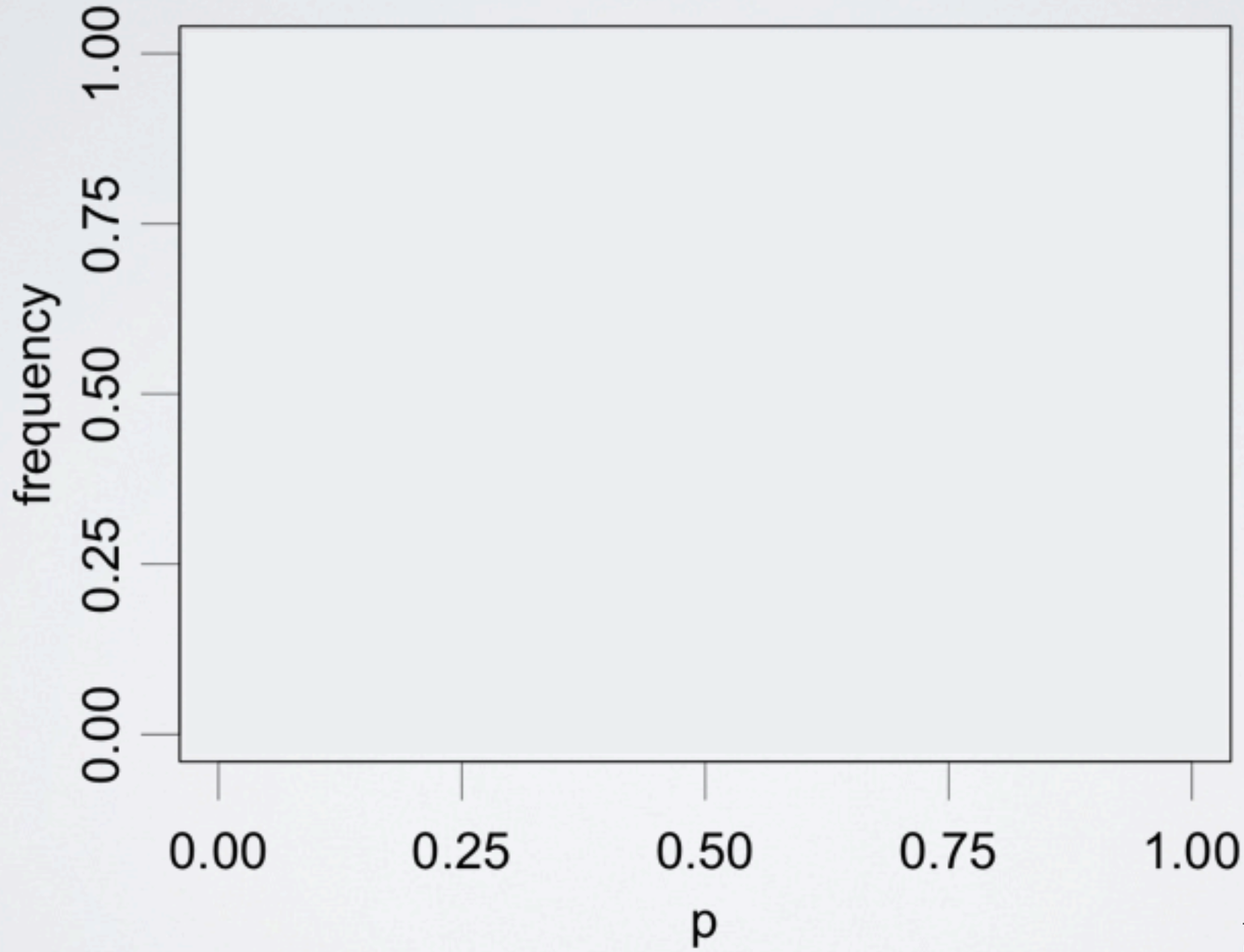
50 node graphs
Orange – Neighborhoods
Green – Groups
Lavender – Events

The triad space

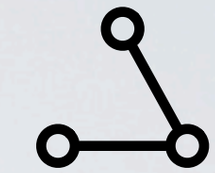


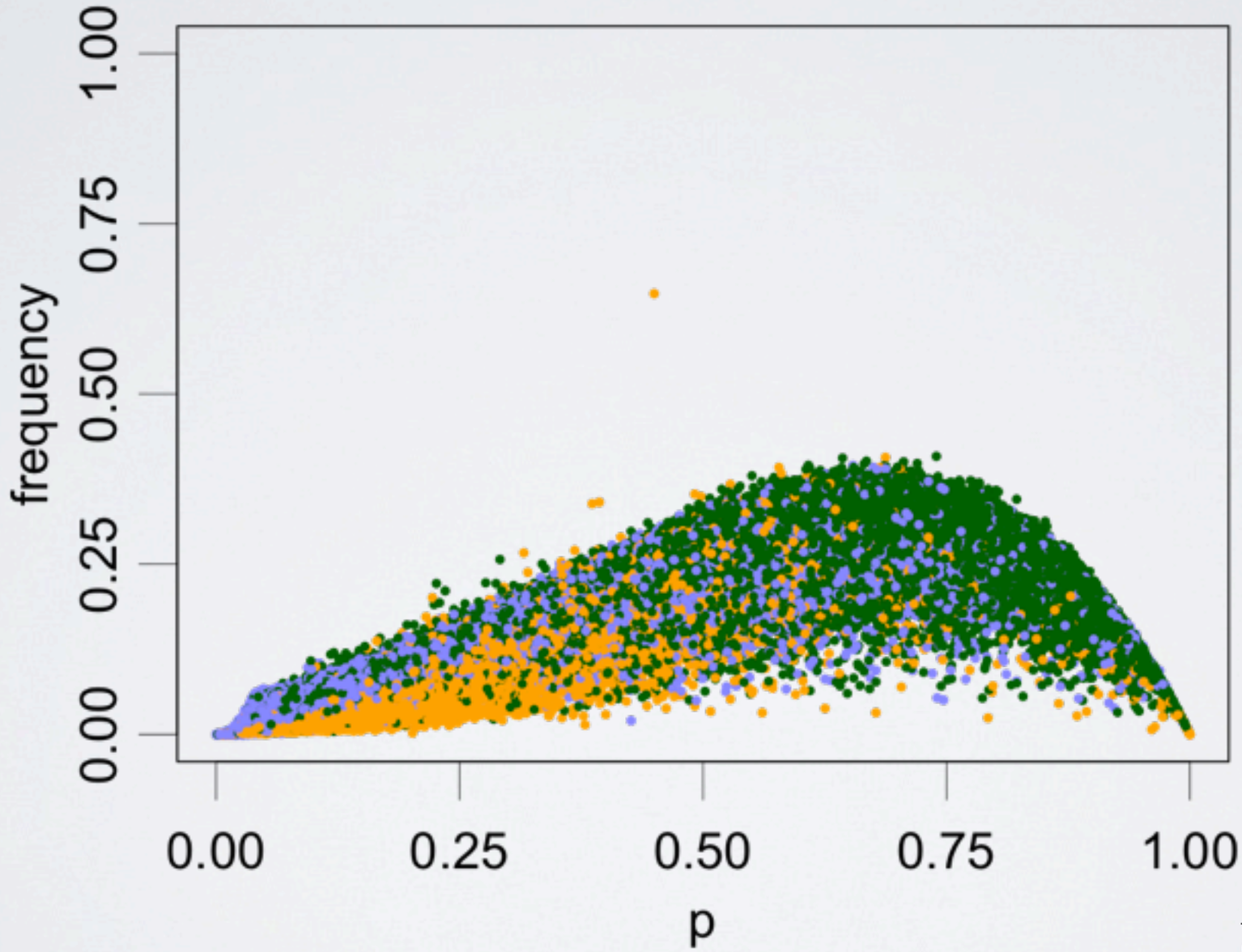
50 node graphs
Orange – Neighborhoods
Green – Groups
Lavender – Events

Subgraph frequency of 

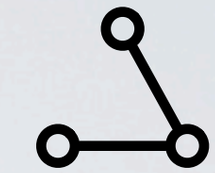


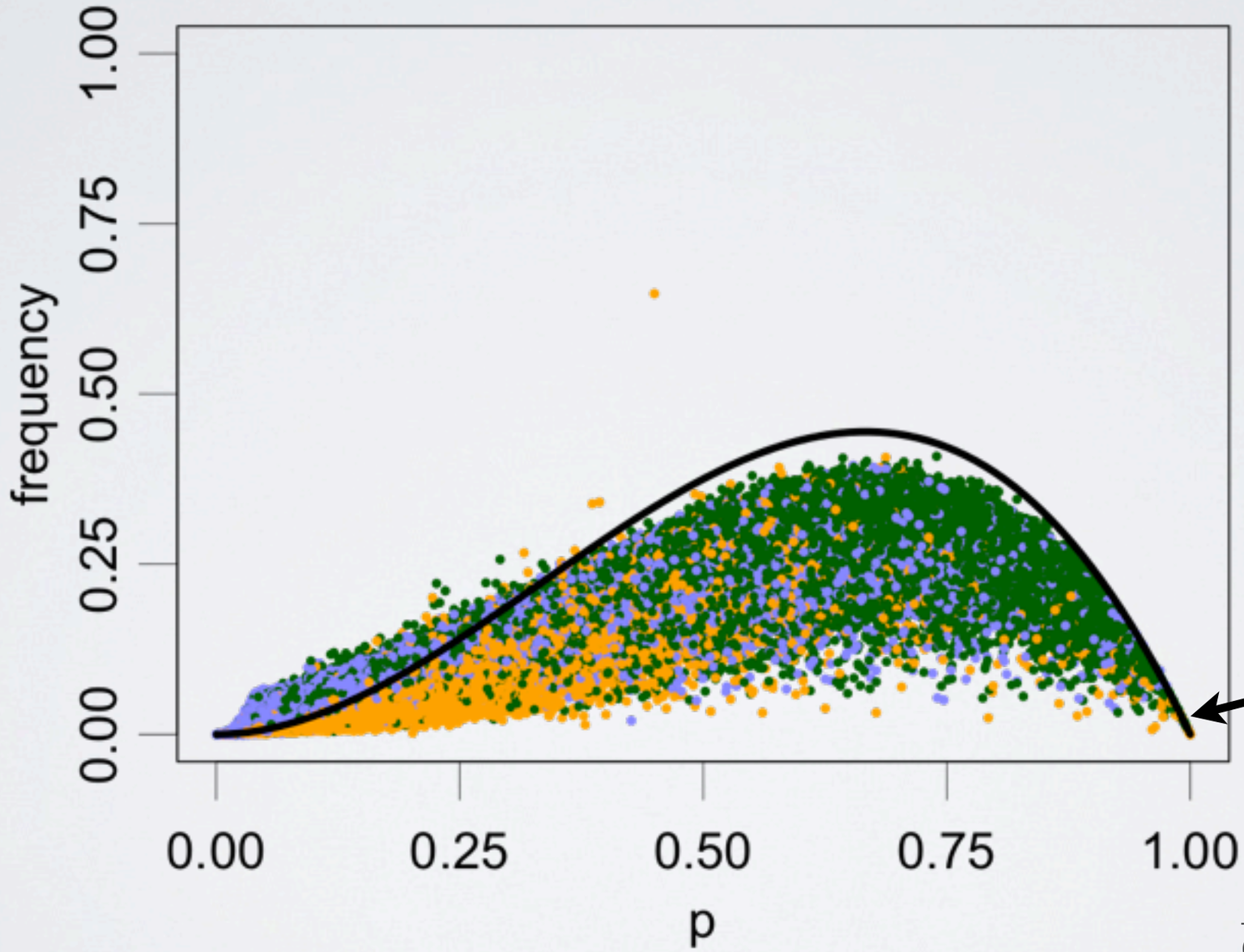
50 node graphs
Orange – Neighborhoods
Green – Groups
Lavender – Events

Subgraph frequency of 



50 node graphs
Orange – Neighborhoods
Green – Groups
Lavender – Events

Subgraph frequency of 



50 node graphs
Orange – Neighborhoods
Green – Groups
Lavender – Events

Moreno: ‘chance sociograms’

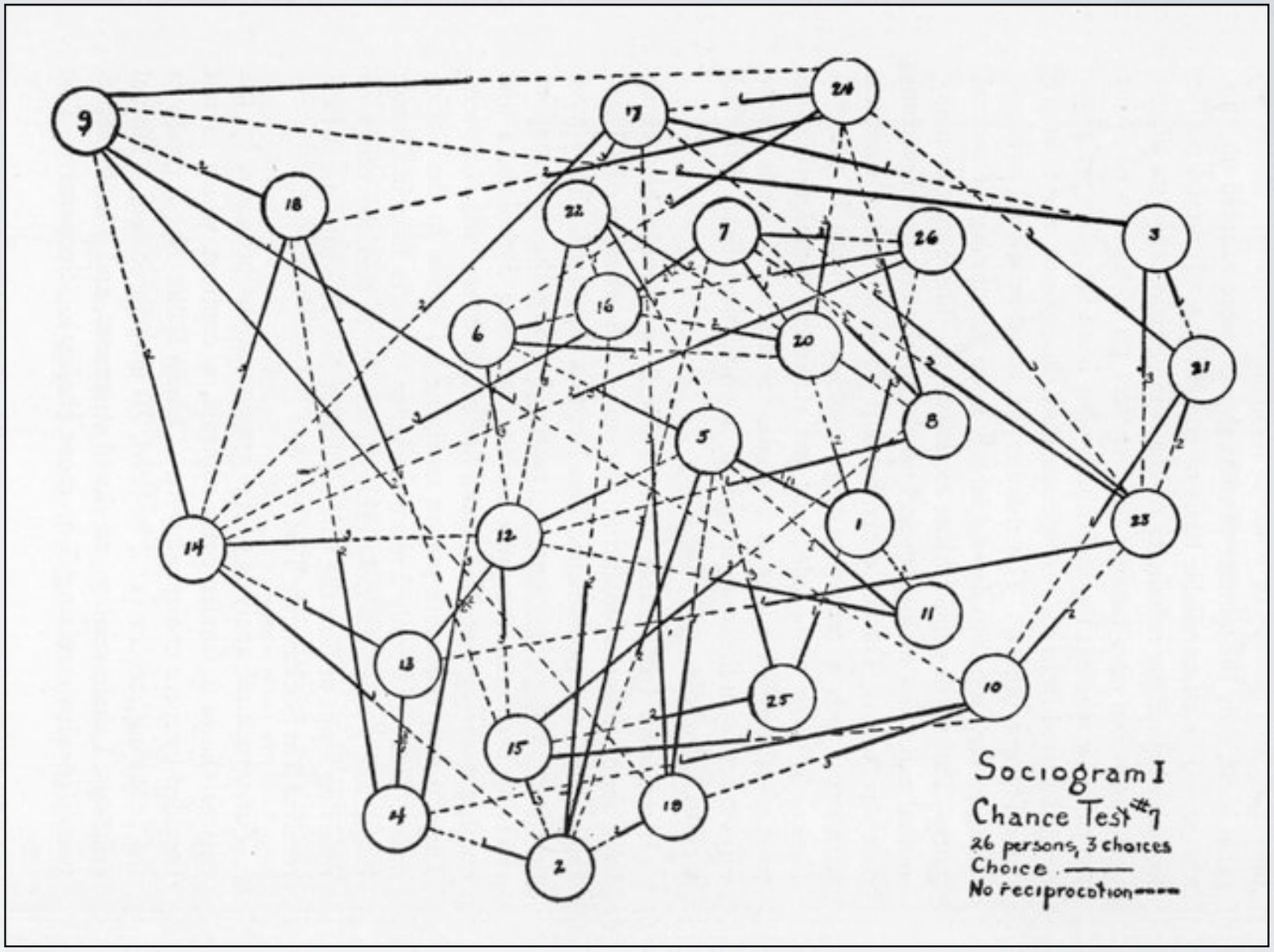
EMOTIONS MAPPED BY NEW GEOGRAPHY

Charts Seek to Portray the Psychological Currents of Human Relationships.

FIRST STUDIES EXHIBITED

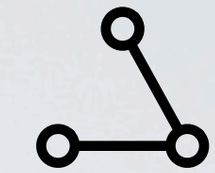
Colored Lines Show Likes and Dislikes of Individuals and of Groups.

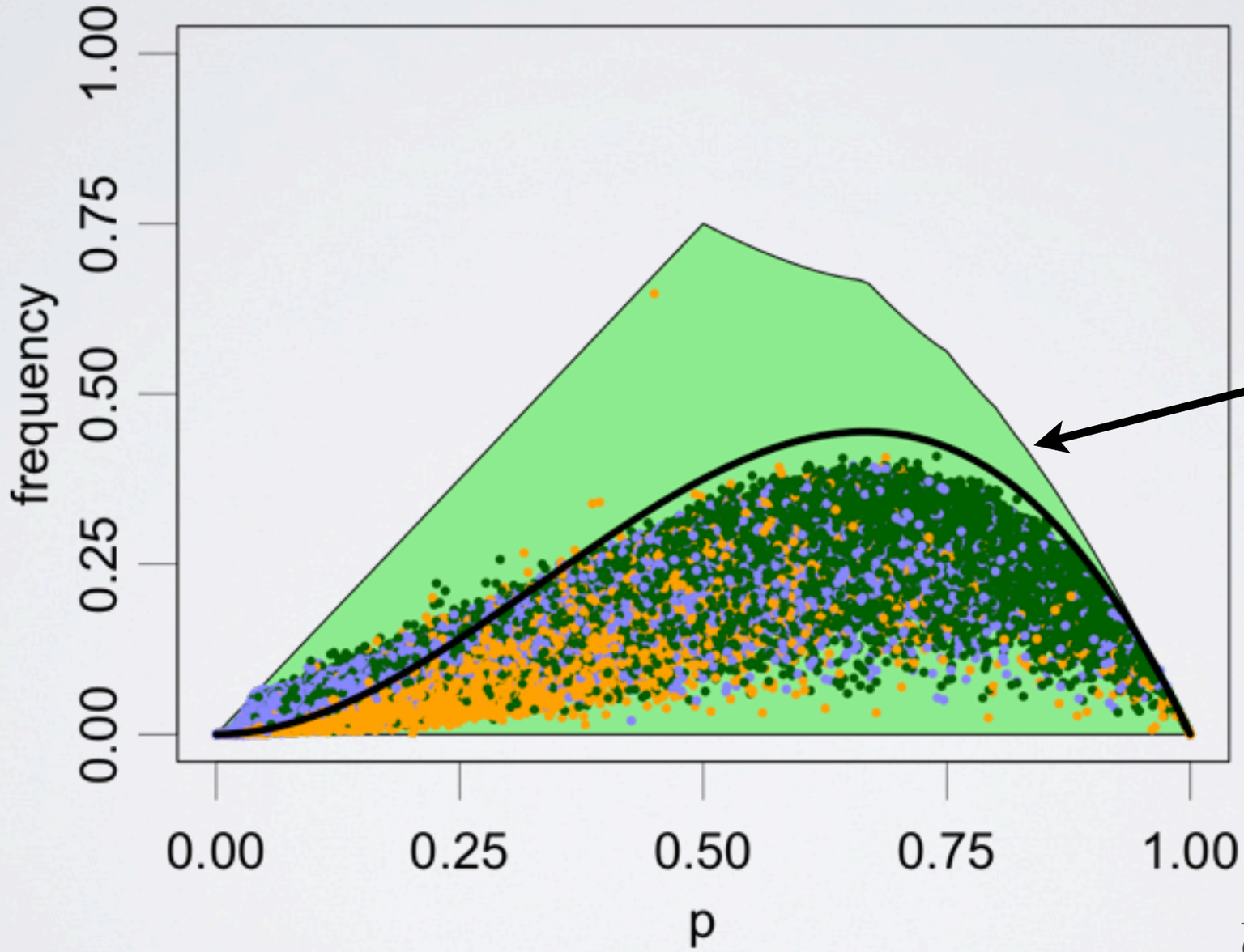
MANY MISFITS REVEALED



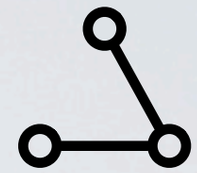
NY Times: April 3, 1933
Presented at 127th Annual meeting of the Medical Society of the State of New York

Moreno, ‘Who Shall Survive?, 1933
Newman and Girvan, PRE, 2004, etc.

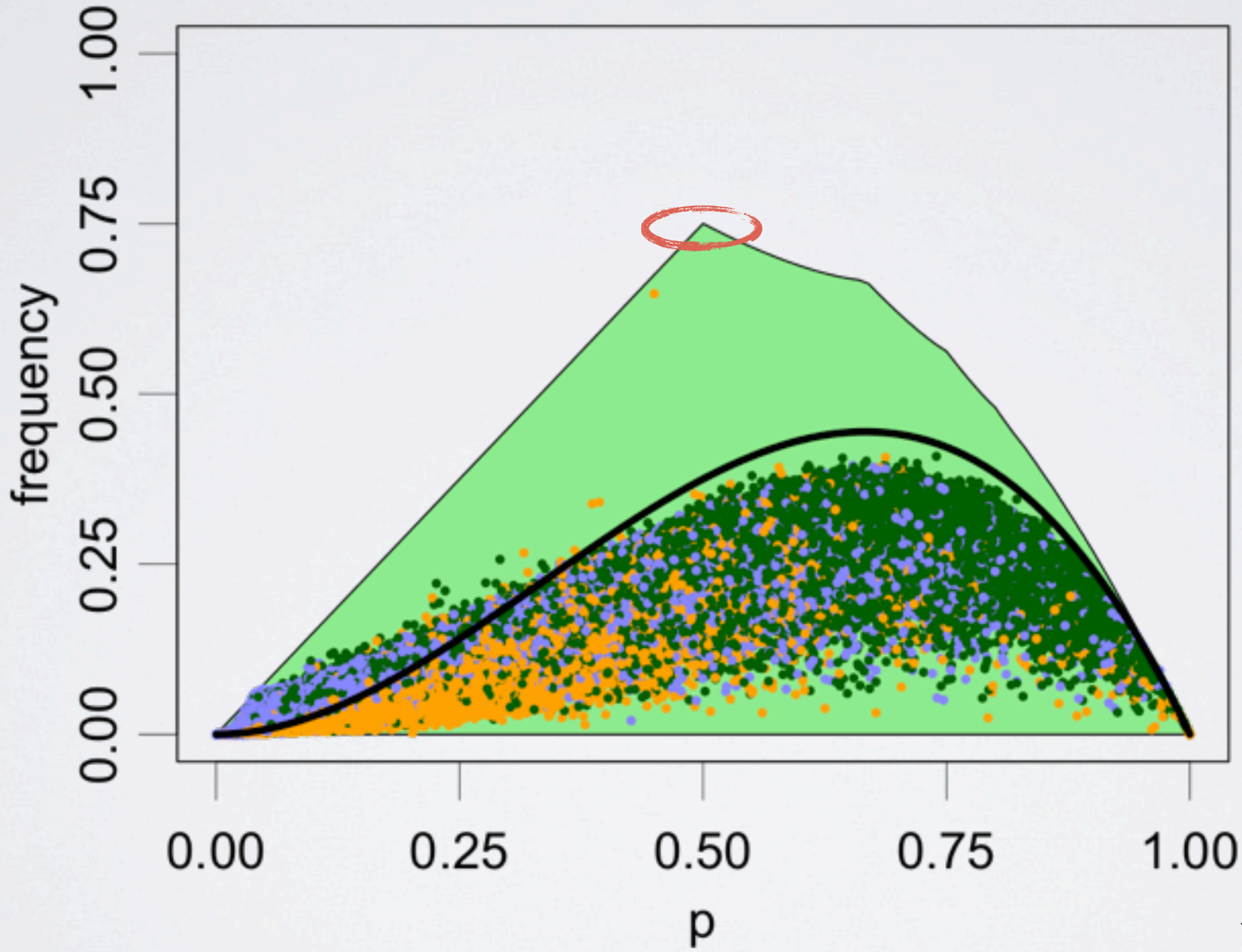
Subgraph frequency of 



50 node graphs
Orange – Neighborhoods
Green – Groups
Lavender – Events

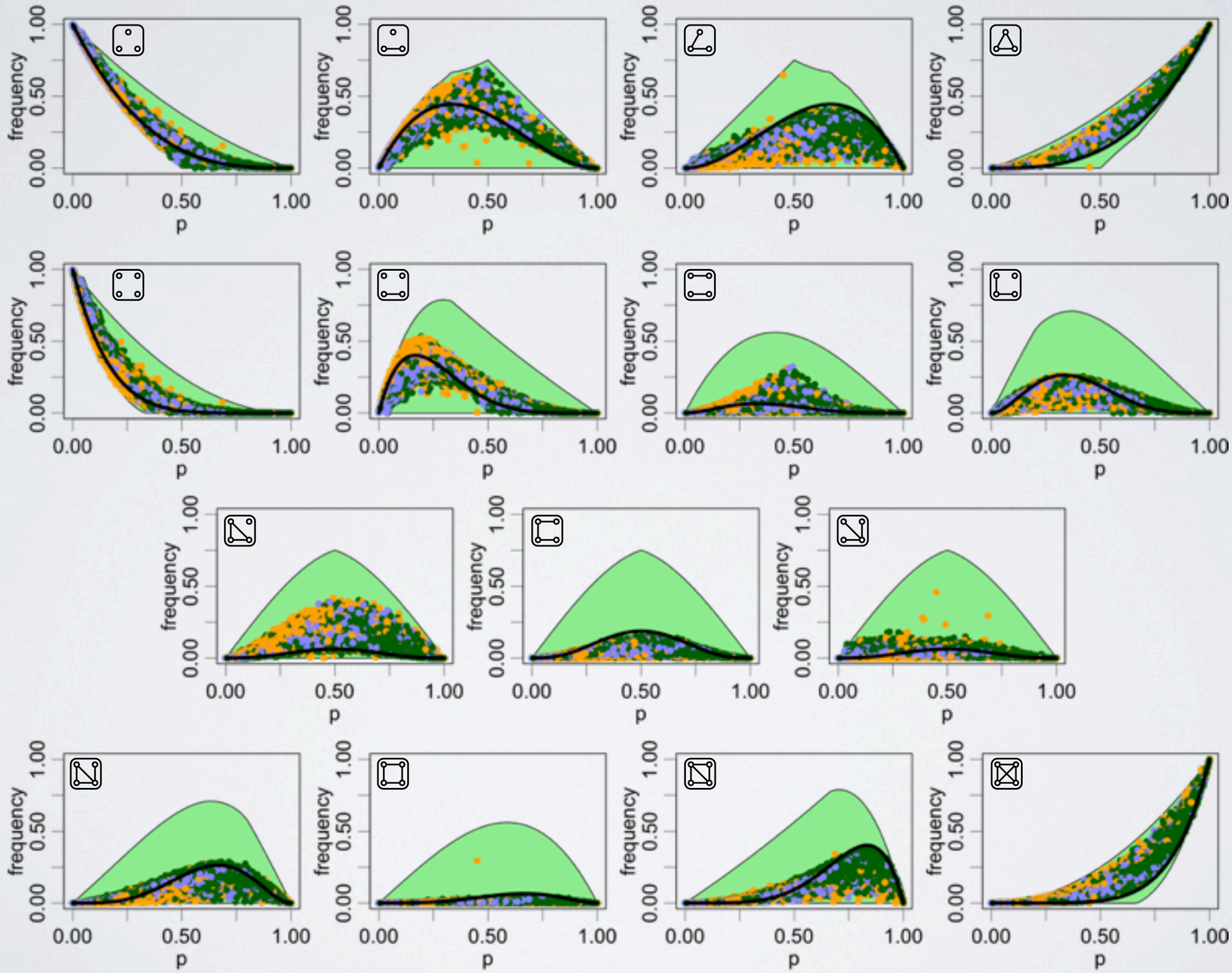
Subgraph frequency of 

Frequency of the ‘forbidden triad’ is bounded at ≤ **3/4** + o(1).
Sharp for $K_{n/2,n/2}$ (bipartite graph) when n is even.

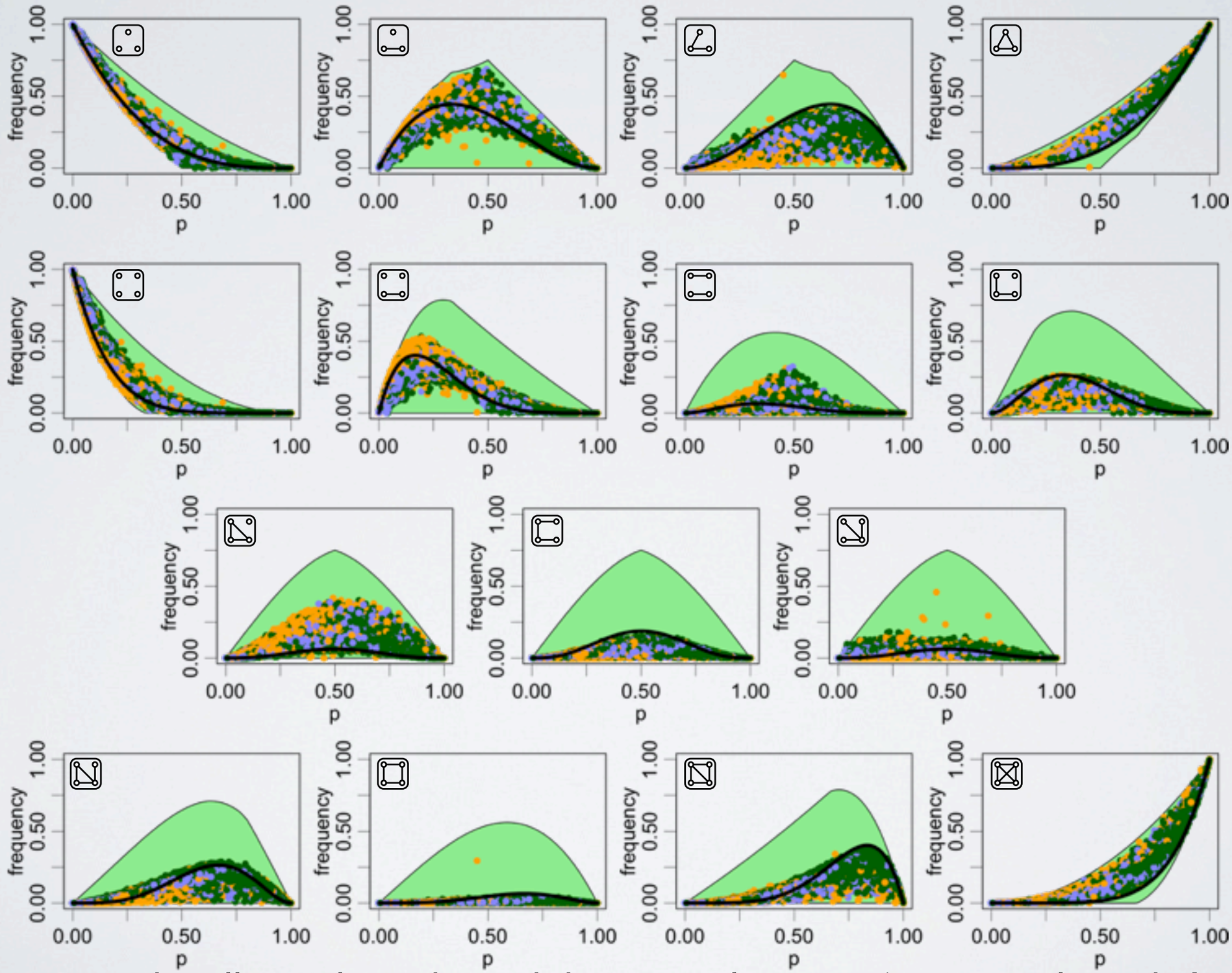


50 node graphs
Orange – Neighborhoods
Green – Groups
Lavender – Events

Subgraph frequencies

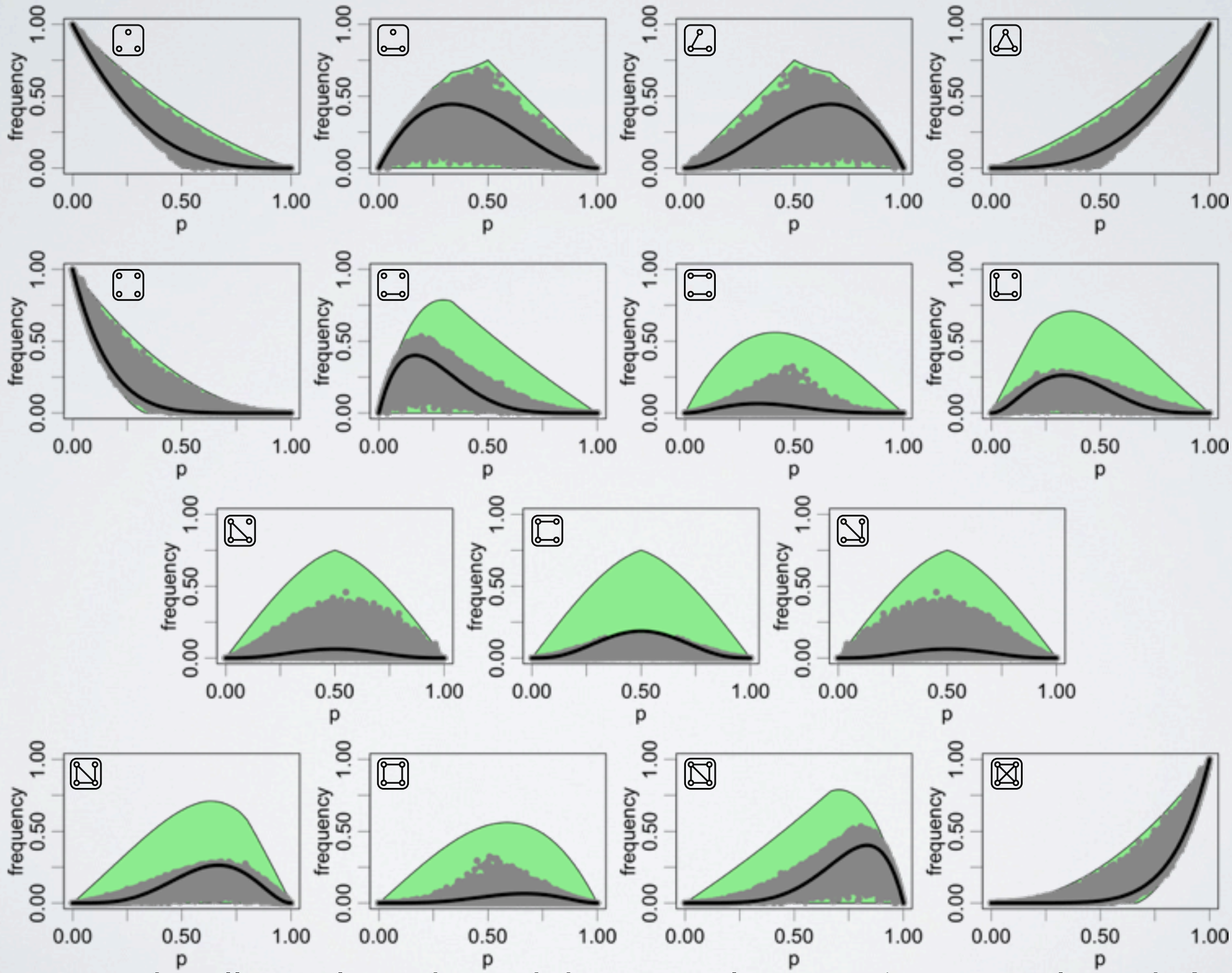


‘Crowd-sourcing’ extremal graphs?



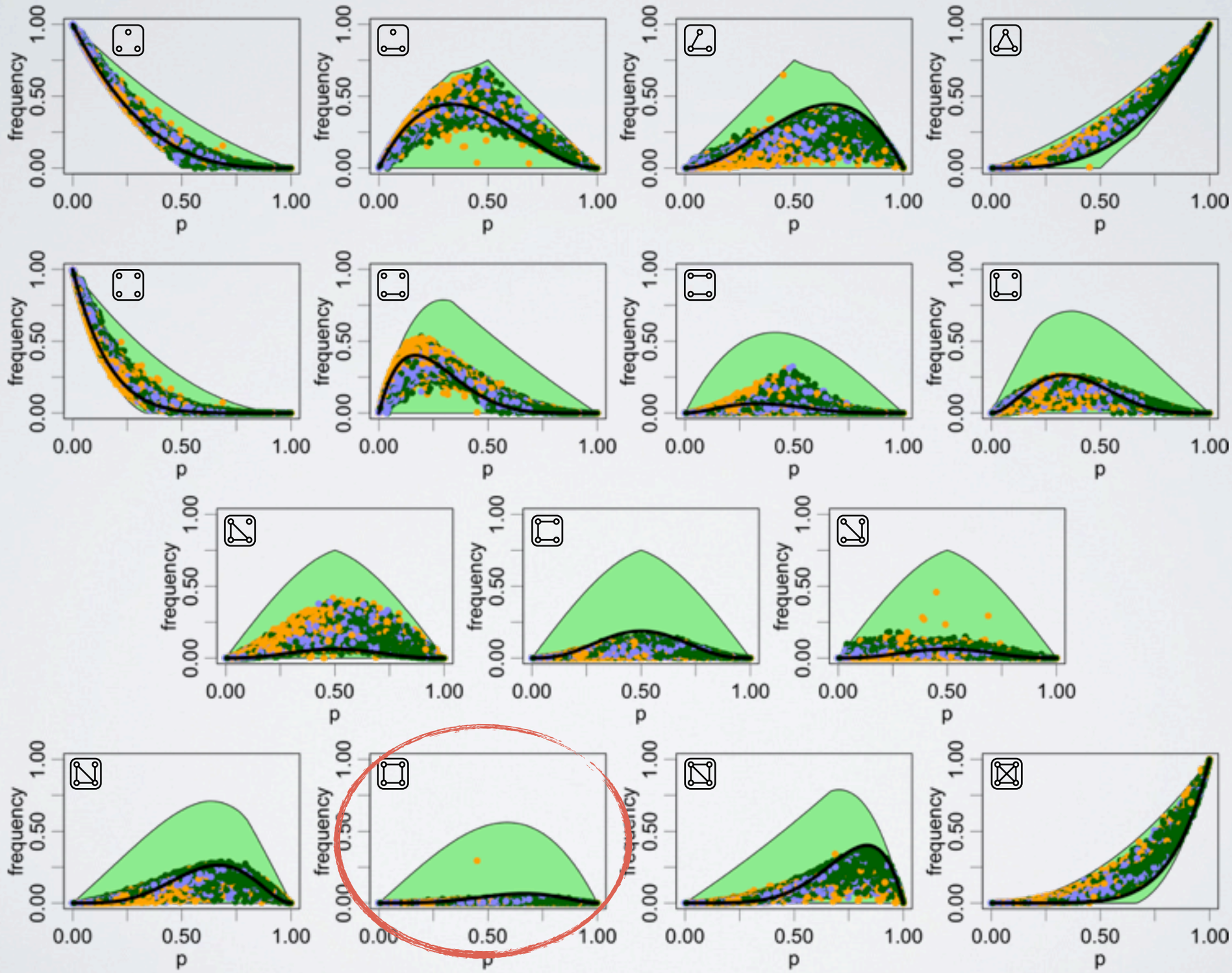
Idea: Consider all social graphs and their complements: ‘anti-social graphs’

‘Crowd-sourcing’ extremal graphs



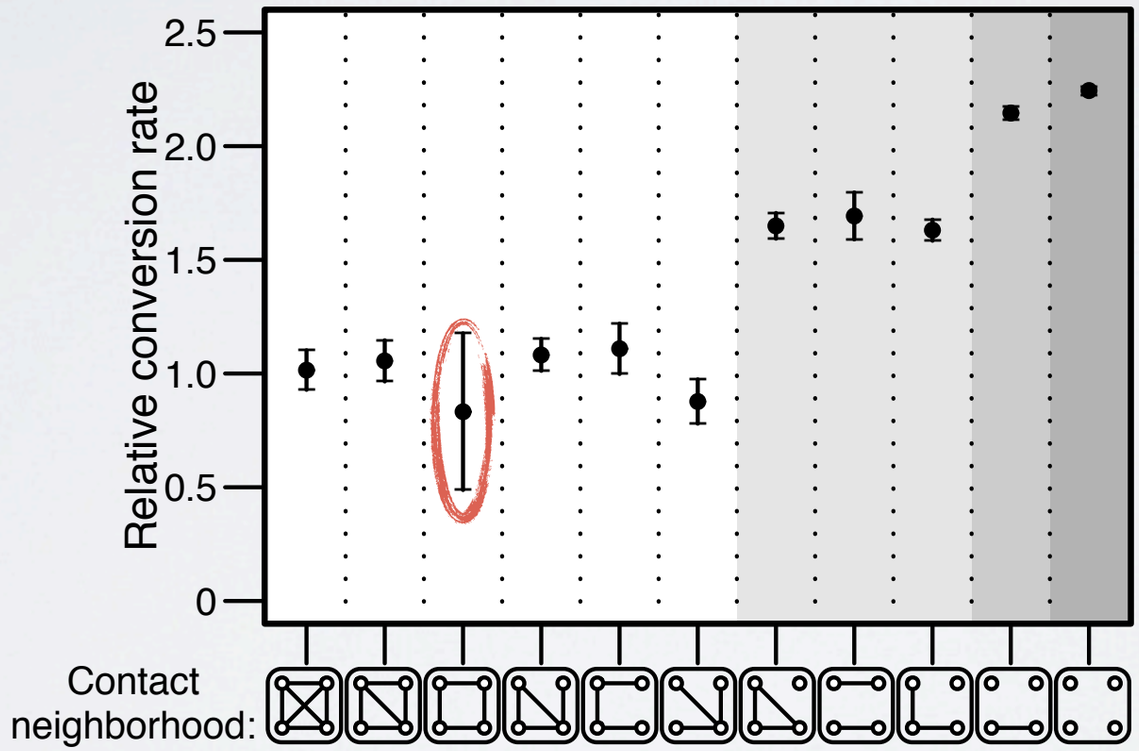
Idea: Consider all social graphs and their complements: ‘anti-social graphs’

What subgraphs are missing?



Squares

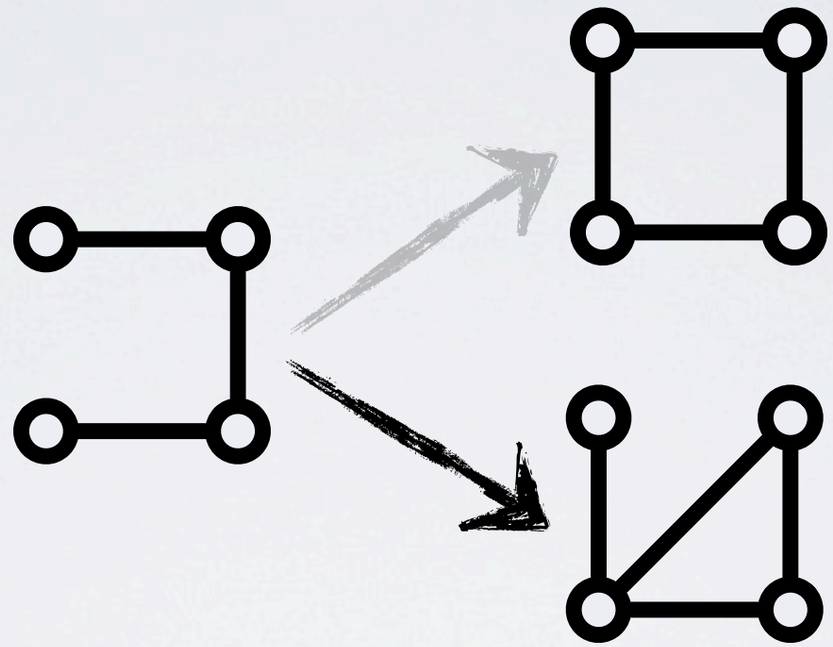
Squares are rare in Facebook neighborhoods:



J. Ugander, L. Backstrom, C. Marlow, J. Kleinberg
Structural Diversity in Social Contagion, PNAS 2012

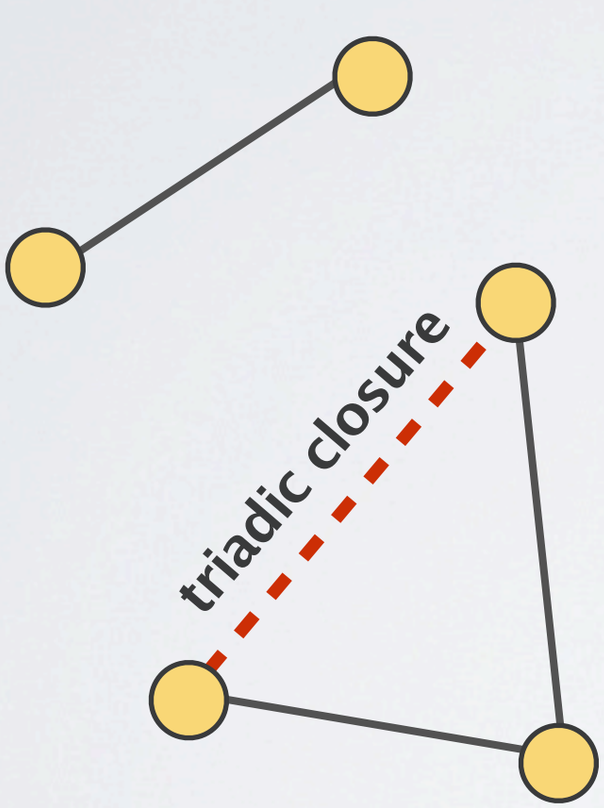
Squares and Triadic Closure

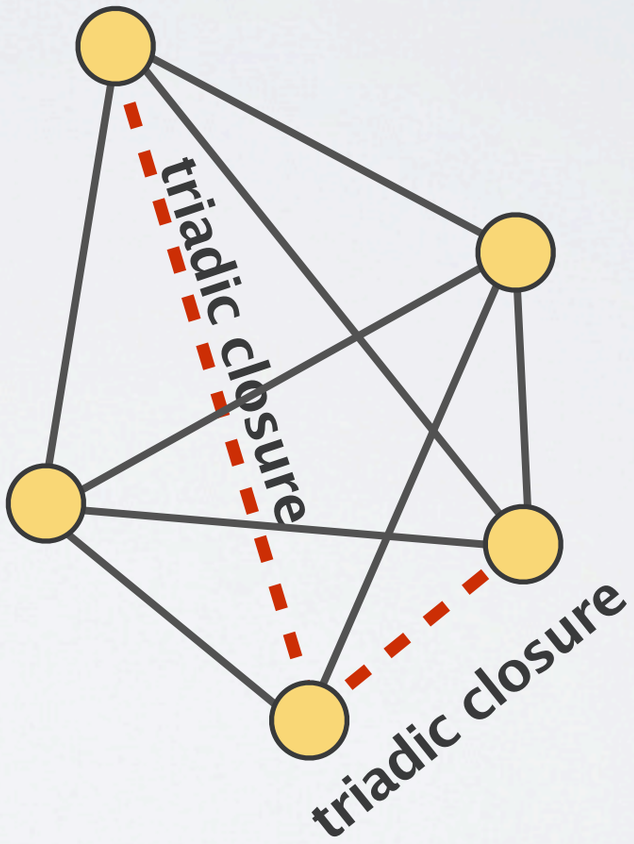
- Square unlikely to form:



- Square has very short '**half-life**':

Continuous Time Markov Chain

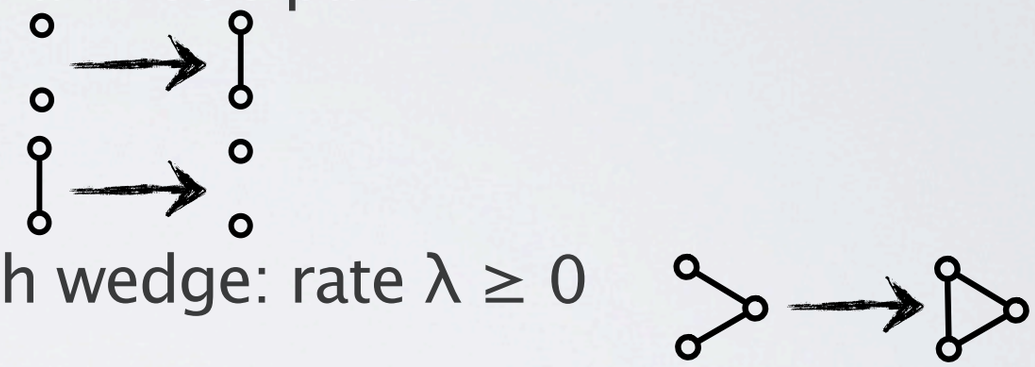




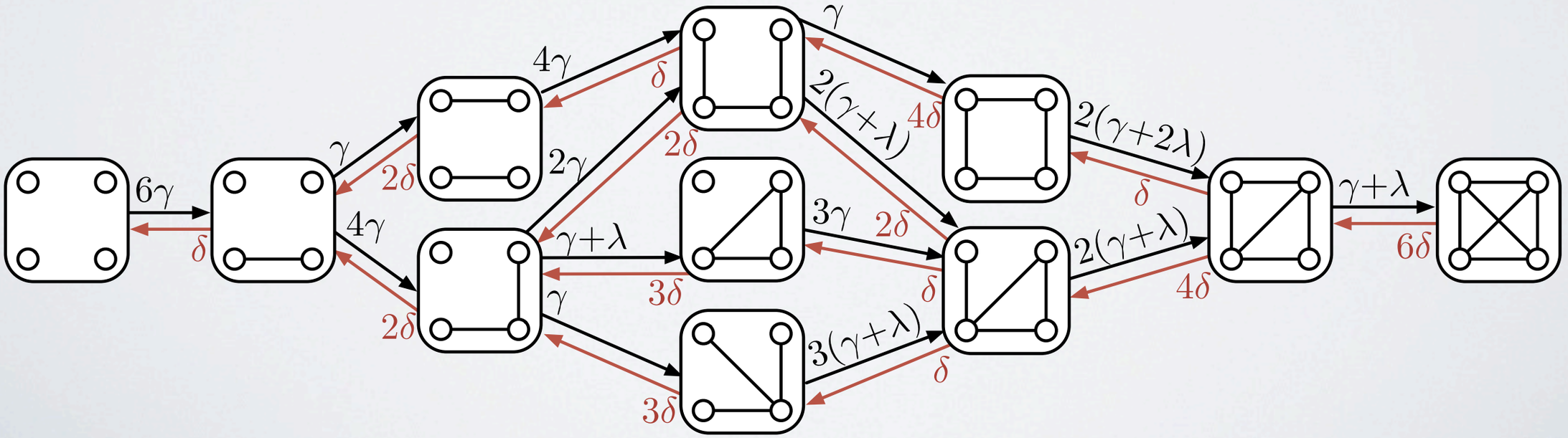
Edge Formation Random Walk

- **Continuous-time Markov chain**
- Transitions between unlabeled, undirected graphs based in edge formation.

- Independent **Poisson processes** for all node pairs:
  - Arbitrary formation: rate $\gamma > 0$
  - Arbitrary deletion: rate $\delta > 0$
  - Triadic closure formation for each wedge: rate $\lambda \geq 0$



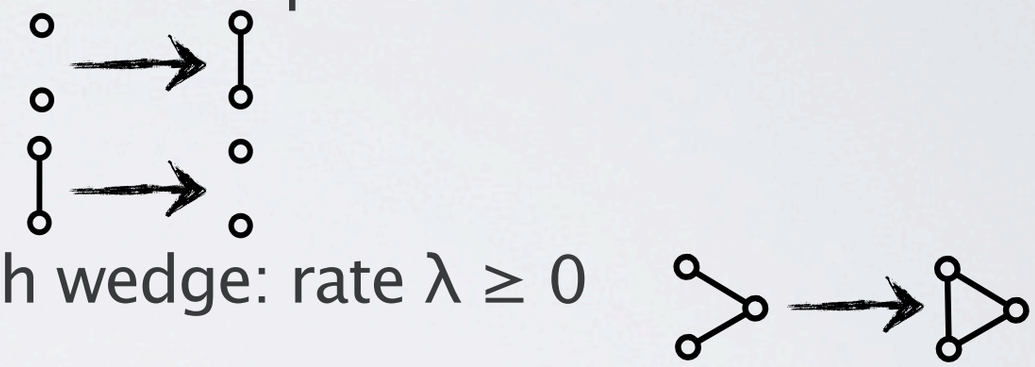
- For 4-node graphs, succinct Markov chain state transition diagram:



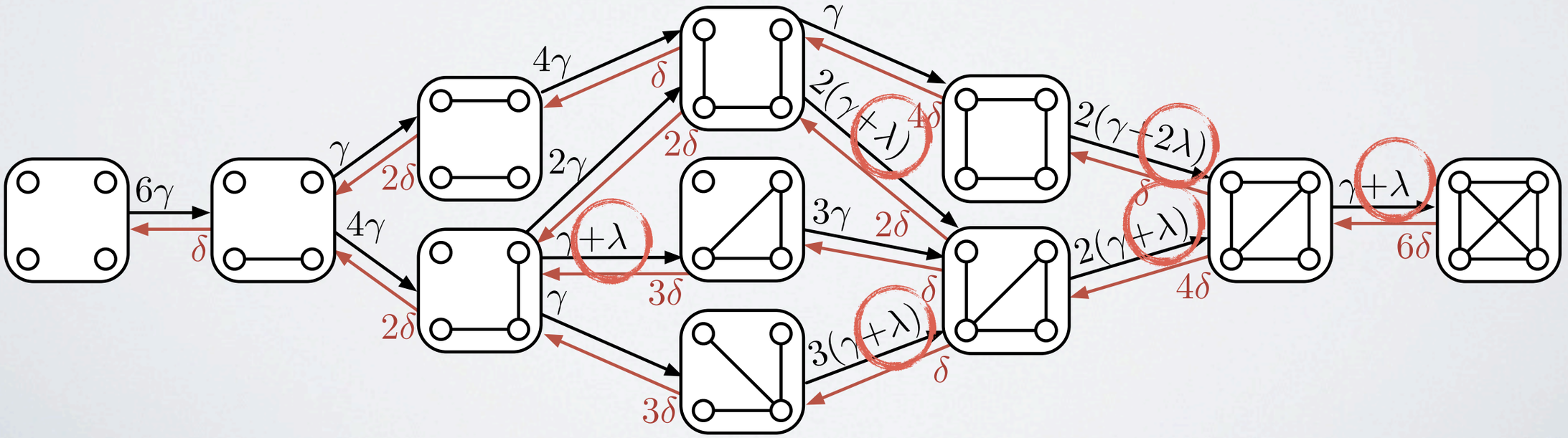
Edge Formation Random Walk

- **Continuous-time Markov chain**
- Transitions between unlabeled, undirected graphs based in edge formation.

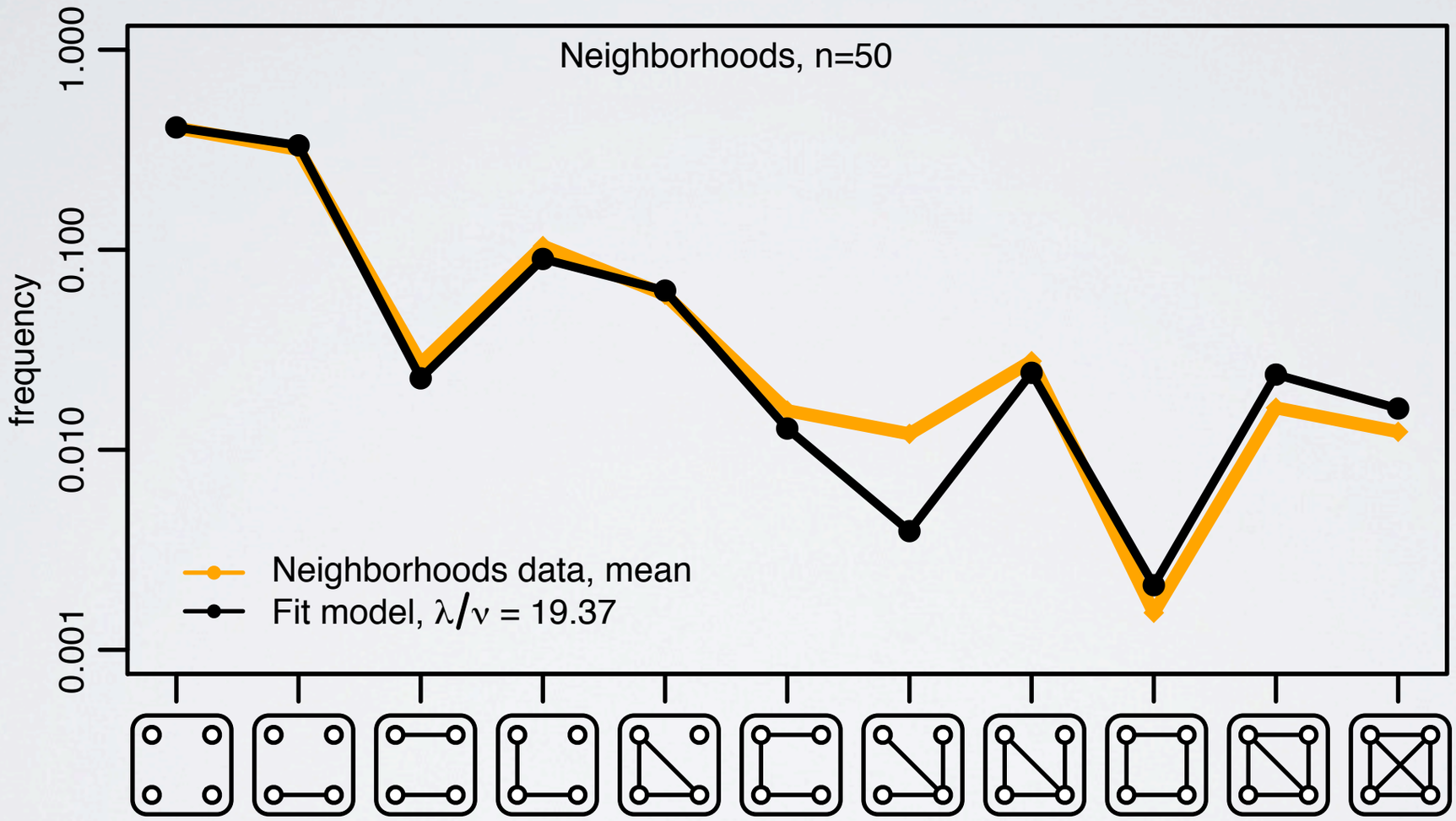
- Independent **Poisson processes** for all node pairs:
  - Arbitrary formation: rate $\gamma > 0$
  - Arbitrary deletion: rate $\delta > 0$
  - Triadic closure formation for each wedge: rate $\lambda \geq 0$



- For 4-node graphs, succinct Markov chain state transition diagram:



Fitting λ to data



(log-scale y-axis)

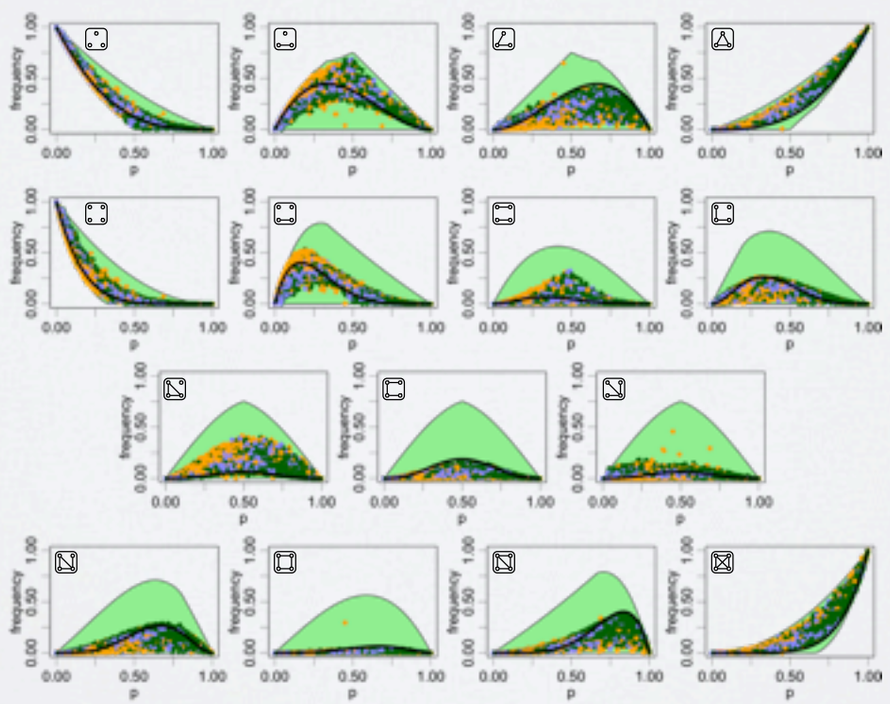
A social model of subgraphs

Extremal graph theory: machinery

- Subgraph frequencies **s(F,G)** relates to homomorphism density **t(F,G)**.
  [Borgs et al. 2006, Lovasz 2009]

- Frequency of cliques, lower bounds: **Moon–Moser** 1962, **Razborov** 2008
- Frequency of cliques, upper bounds: **Kruskal–Katona Theorem**
- Frequency of trees: **Sidorenko Conjecture** (Theorem for trees)
- Also linear relationships across sizes.

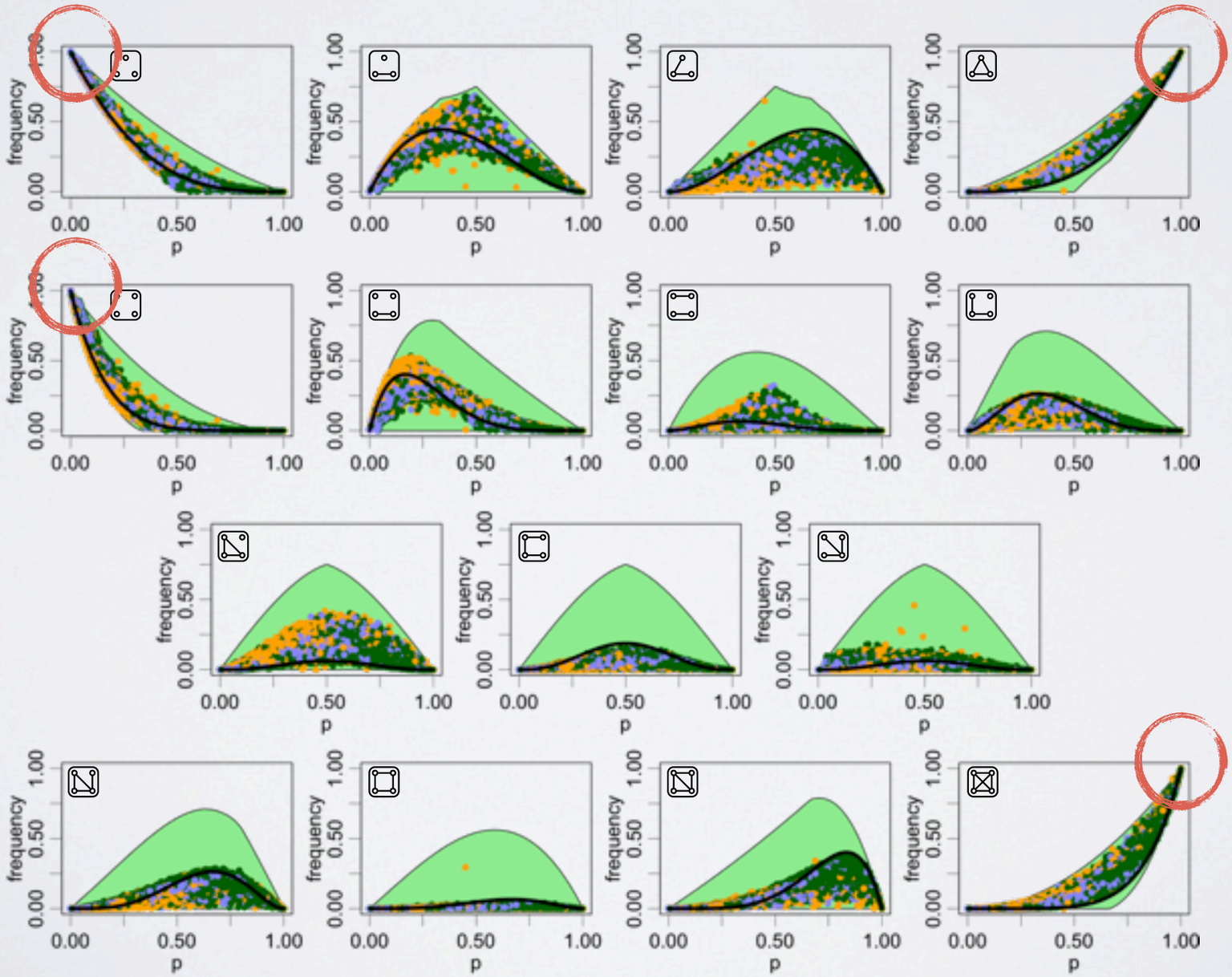
- For each fixed edge density **p,** build a **Linear Program.**



Extremal graph theory: generally

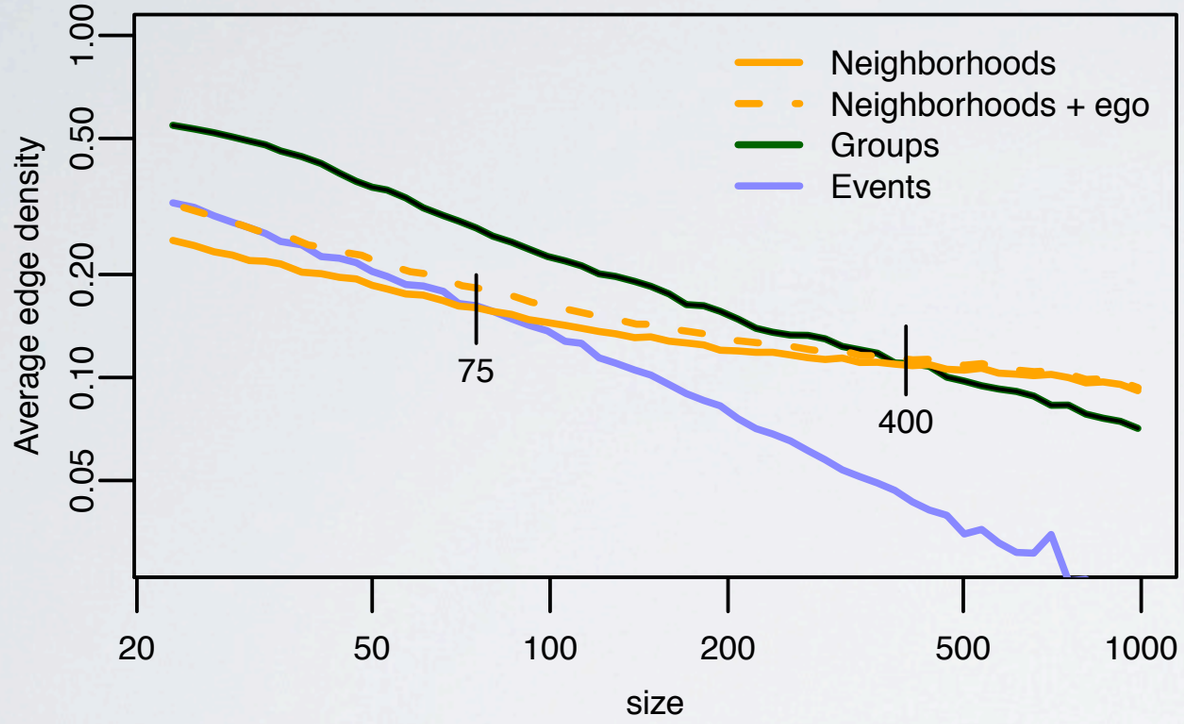
- A proposition for all subgraphs:

Proposition. For every k, there exist constants ϵ and n_0 such that the following holds. If F is a k-node subgraph that is not a clique and not empty, and G is any graph on $n \geq n_0$ nodes, then $s(F, G) < 1 - \epsilon$.



Audience graph classification

- How do different audience graphs differ?

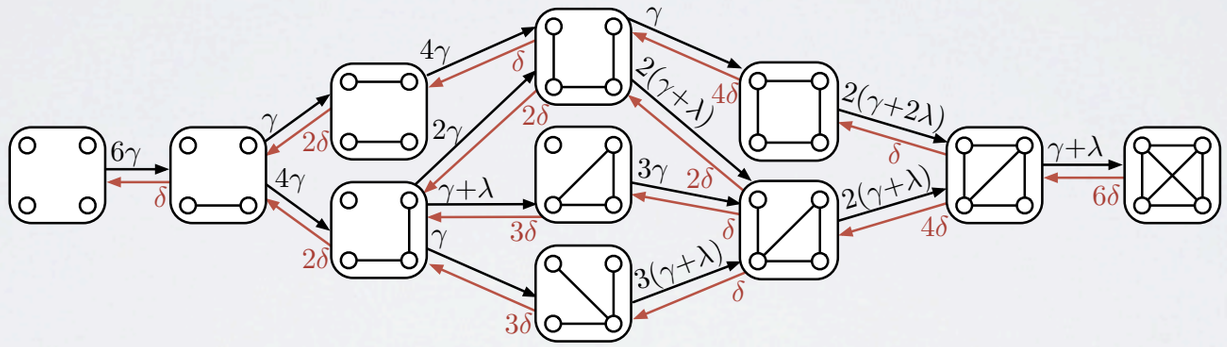


75 nodes: n'hoods v. events
400 nodes: n'hoods vs. groups

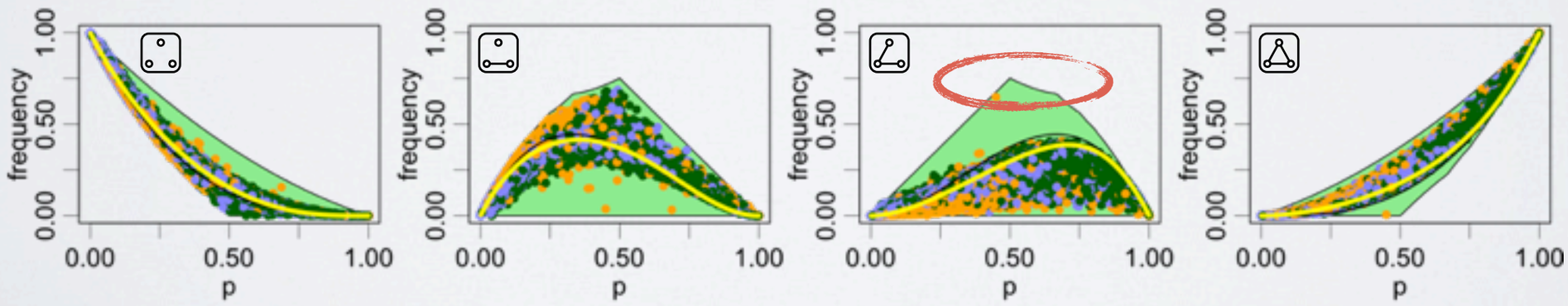
		N vs E	/	N vs G	
Features:	Quad frequencies:	76%	/	76%	accuracy
	Global features:	69%	/	76%	accuracy
	Quad frequencies + Global features:	81%	/	82%	accuracy

Conclusions

- **Combinatorics:** contain non-trivial extremal limits! What else?
- **Study collections:** corpus of graphs, etc. (Goel et al. 2013, Cheng 2014)
- **Edge Formation Random Walk** model of dense social graphs:



- **Homomorphism density bounds** yield subgraph density bounds:



- **Paper, optimization code:** http://cam.cornell.edu/~jugander/