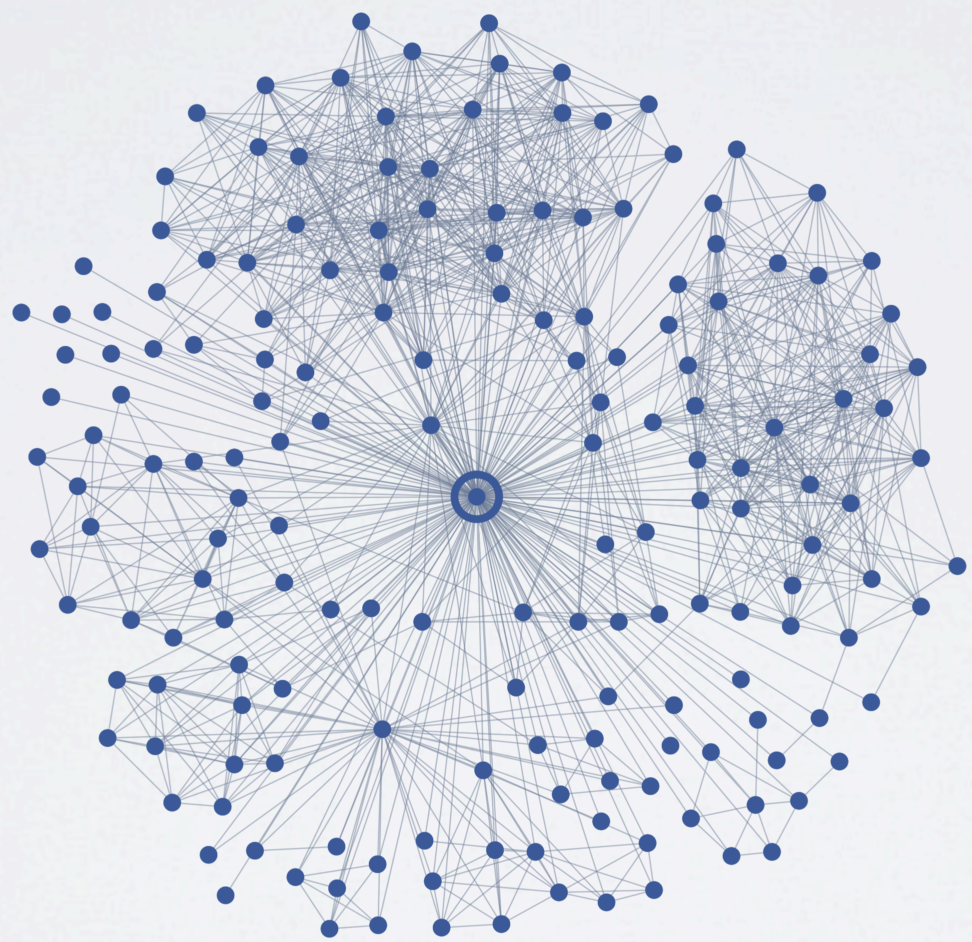
Subgraph Frequencies:
The Empirical and Extremal Geography of Large Graph Collections

Johan Ugander, Lars Backstrom, Jon Kleinberg
World Wide Web Conference
May 16, 2013

Cornell University

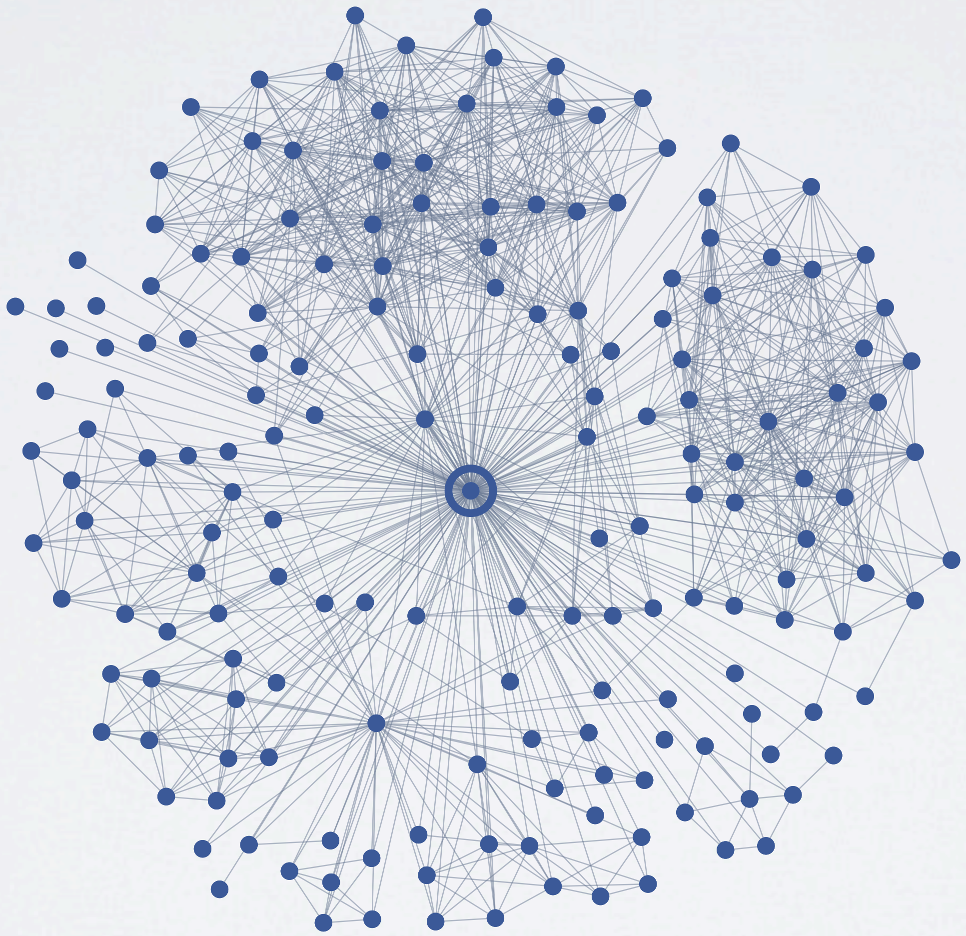
Graph collections

- **Neighborhoods:** graph induced by friends of a single ego, excluding ego



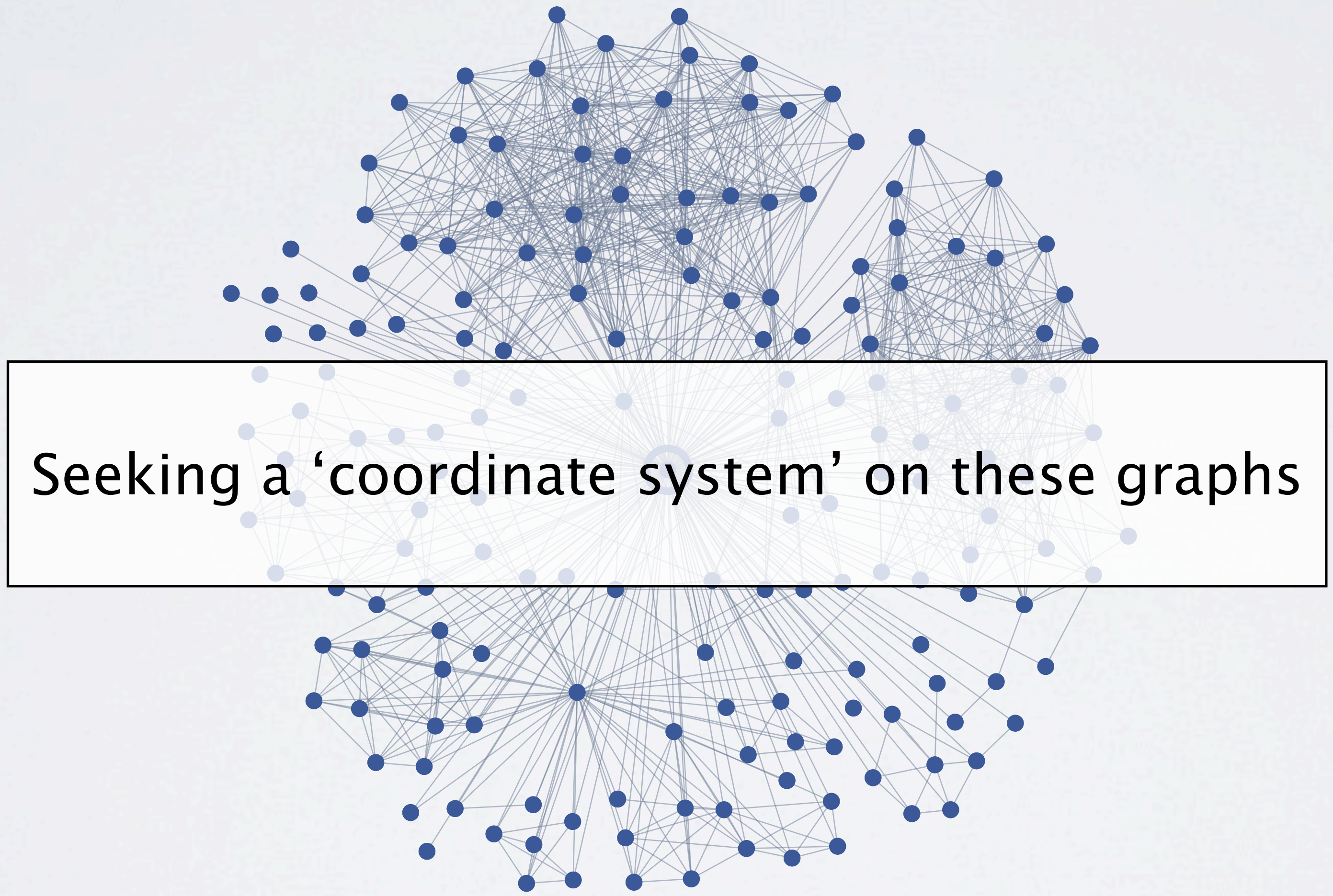
Graph collections

- **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’

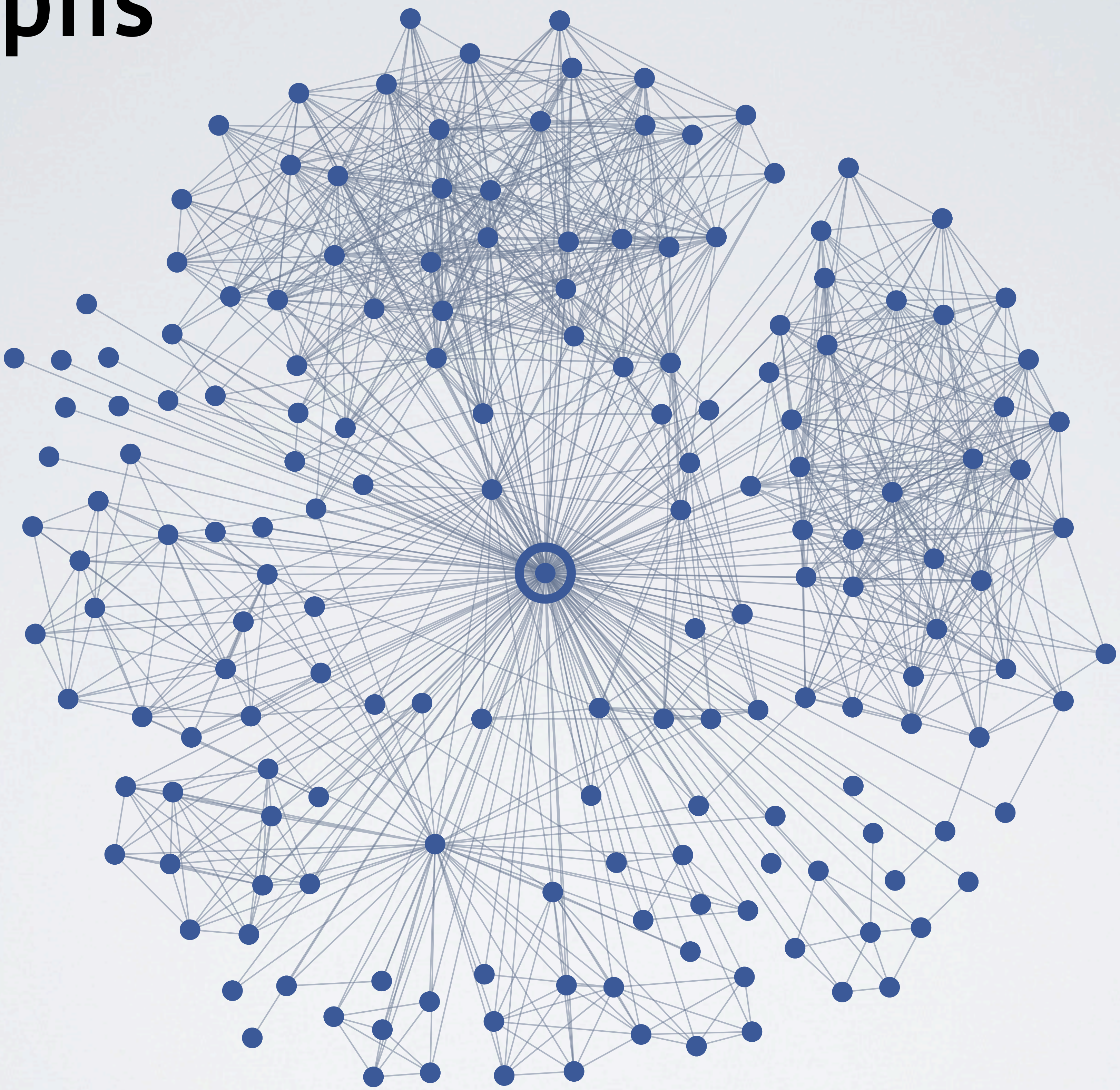


Graph collections

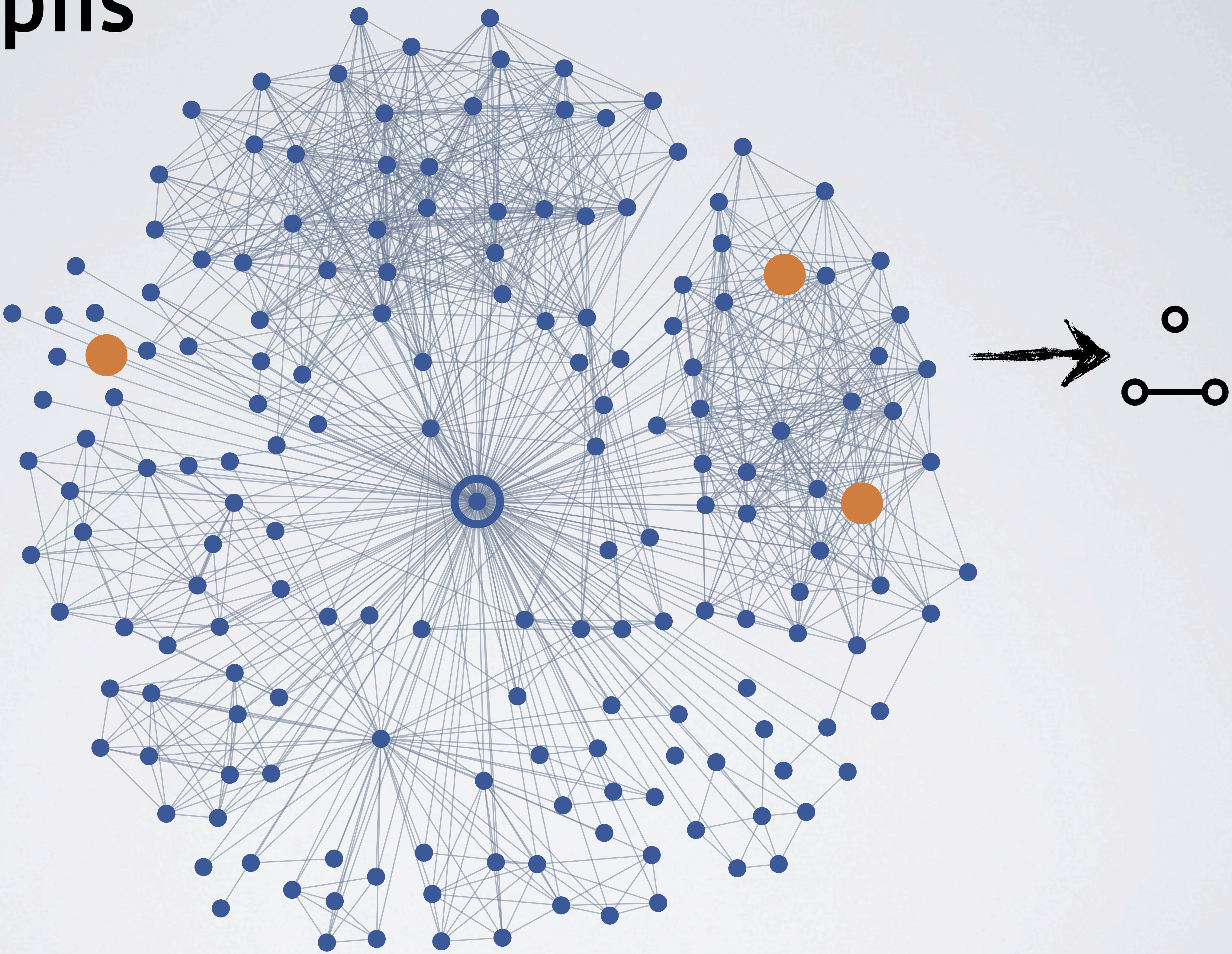
- **Neighborhoods:** graph induced by friends of a single ego, excluding ego
- **Groups:** graph induced by members of a Facebook ‘group‘
- **Events:** graph induced by ‘Yes’ respondents to a Facebook ‘event’



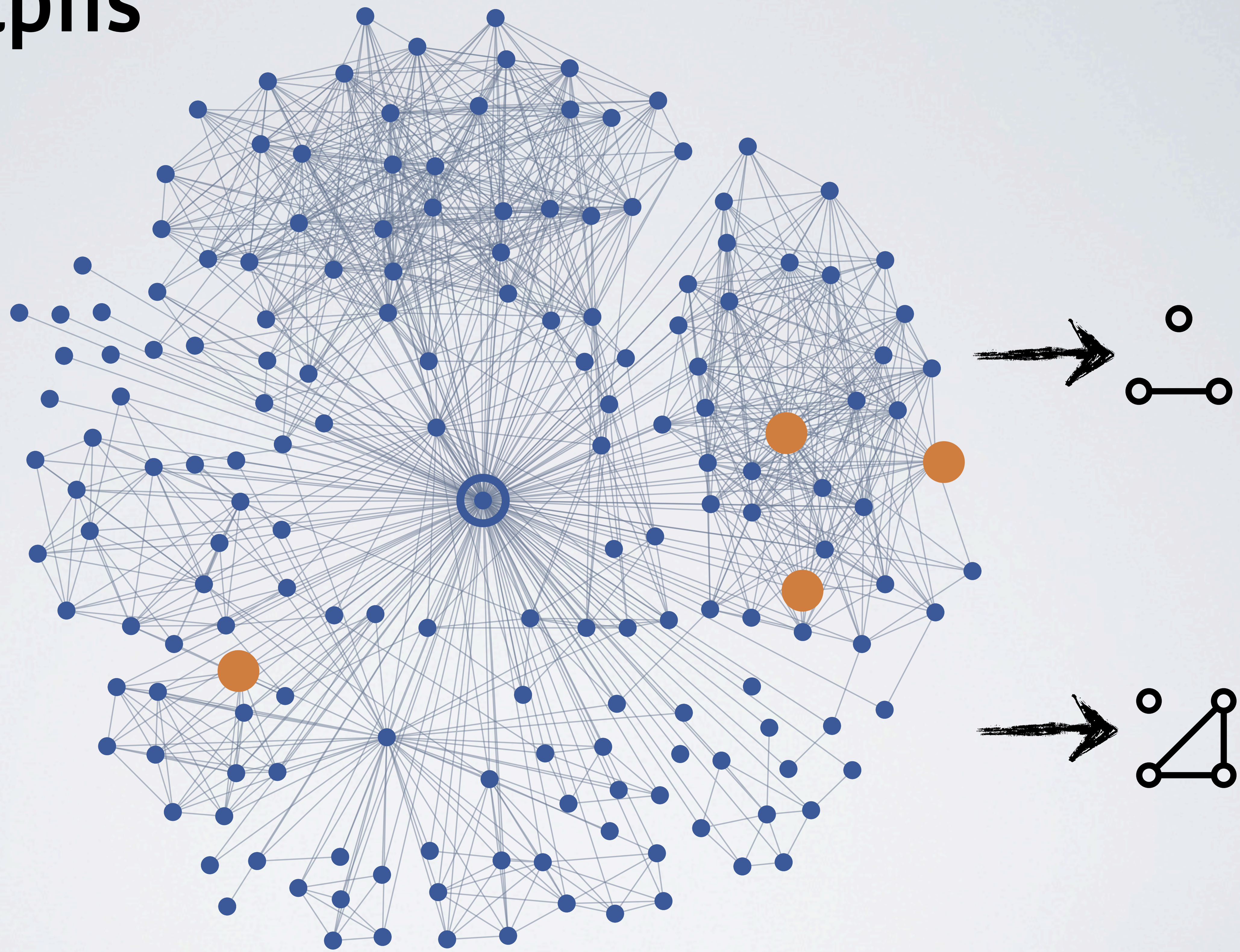
Subgraphs



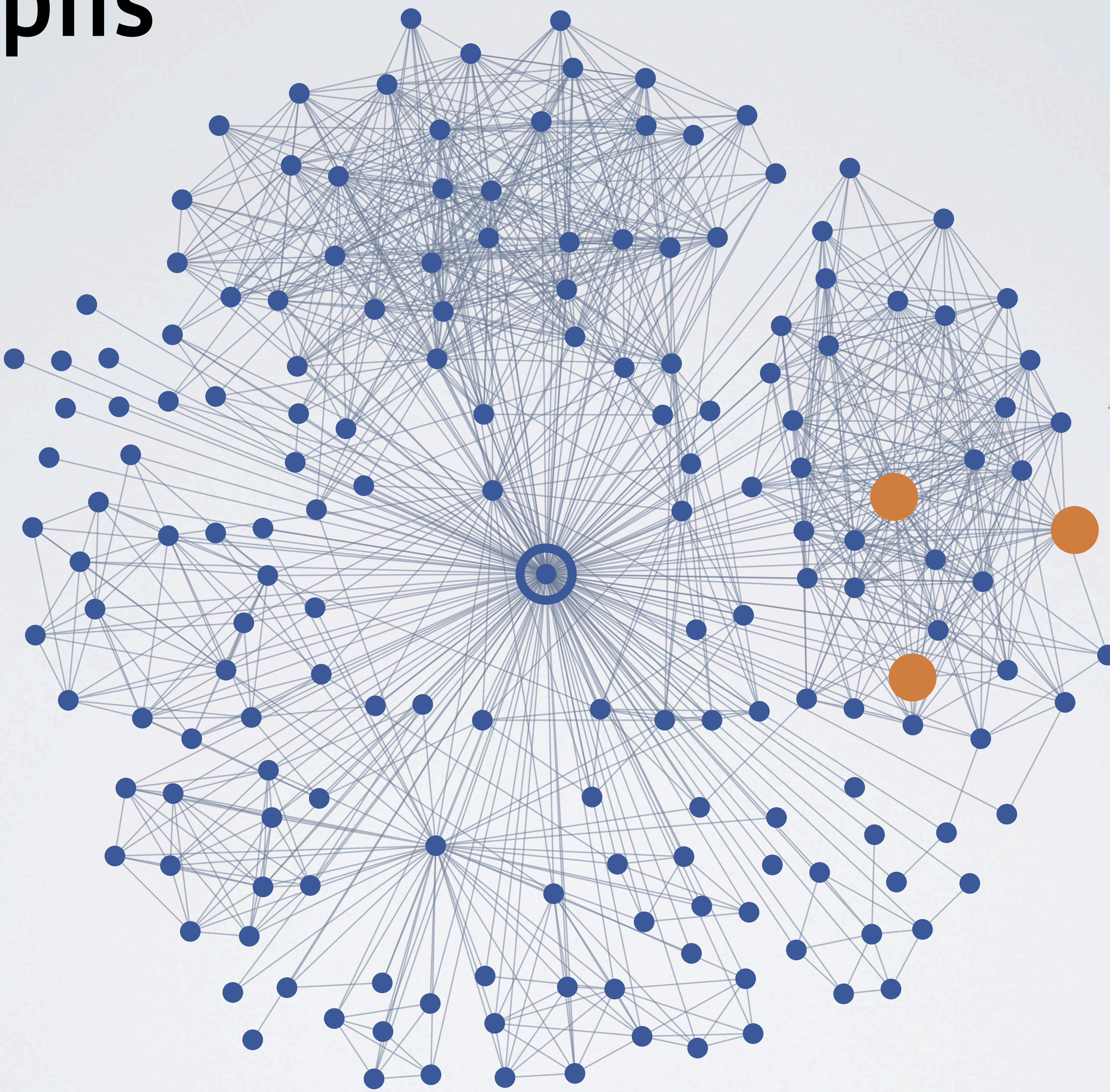
Subgraphs



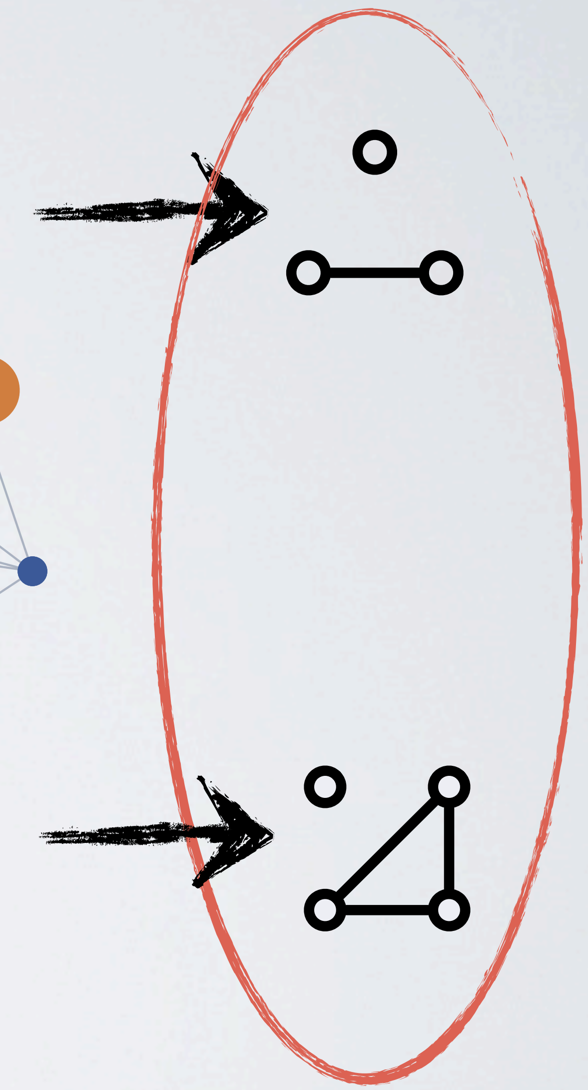
Subgraphs



Subgraphs



Compute frequencies



Subgraph Frequencies

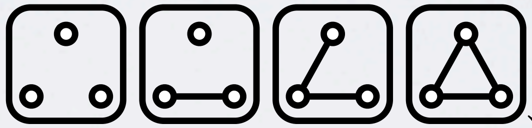
- **Definition:** The **subgraph frequency s(F,G)** of a k-node subgraph **F** in a graph **G** is the fraction of k-tuples of nodes in **G** that induce a copy of **F.**

Triad census: Davis-Leinhardt 1971, Wasserman-Faust 1994
Motifs/Frequent subgraphs: Inokuchi et al. 2000, Milo et al. 2002, Yan-Han 2002, Kuramochi-Karypis 2004

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:**



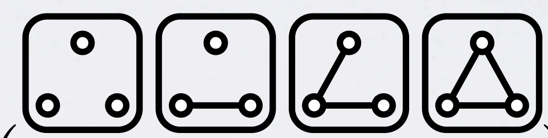
$$s(\cdot, G) = (x_1,\ x_2,\ x_3,\ x_4) = (0.18, 0.37, 0.14, 0.31)$$

Triad census: Davis-Leinhardt 1971, Wasserman-Faust 1994
Motifs/Frequent subgraphs: Inokuchi et al. 2000, Milo et al. 2002, Yan-Han 2002, Kuramochi-Karypis 2004

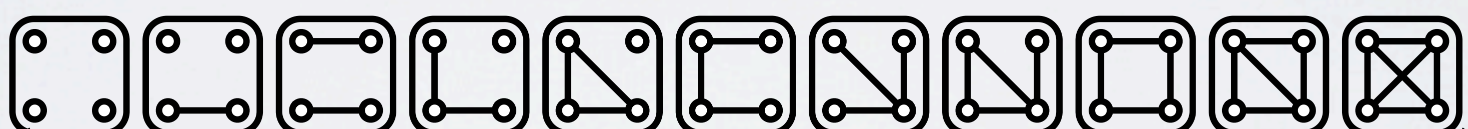
Subgraph Frequencies

- **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, Yan-Han 2002, Kuramochi-Karypis 2004

Empirical/Extremal Questions

- Consider the subgraph frequencies as a ‘**coordinate system**’

- **Empirical Geography:**
  - What subgraph frequencies do **social graphs** exhibit?
  - Is there a good model?

- **Extremal Geography:**
  - How much of this space is even feasible, **combinatorially**?
  - Do empirical graphs fill the **feasible space**?

Empirical/Extremal Questions

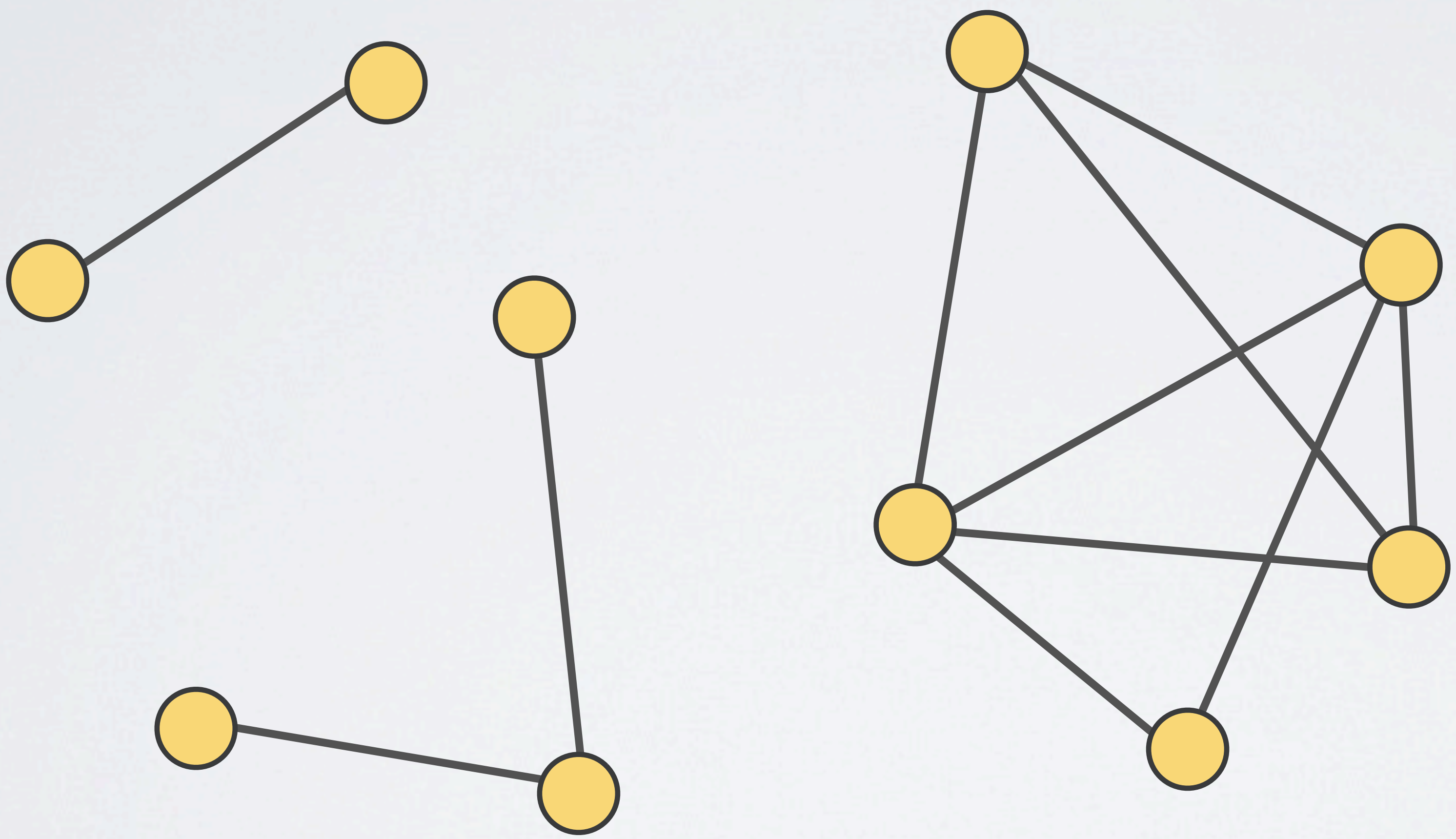
- Consider the subgraph frequencies as a '**coordinate system**'

- **Empirical Geography:**
  - What subgraph frequencies do **social graphs** exhibit?
  - Is there a good model?

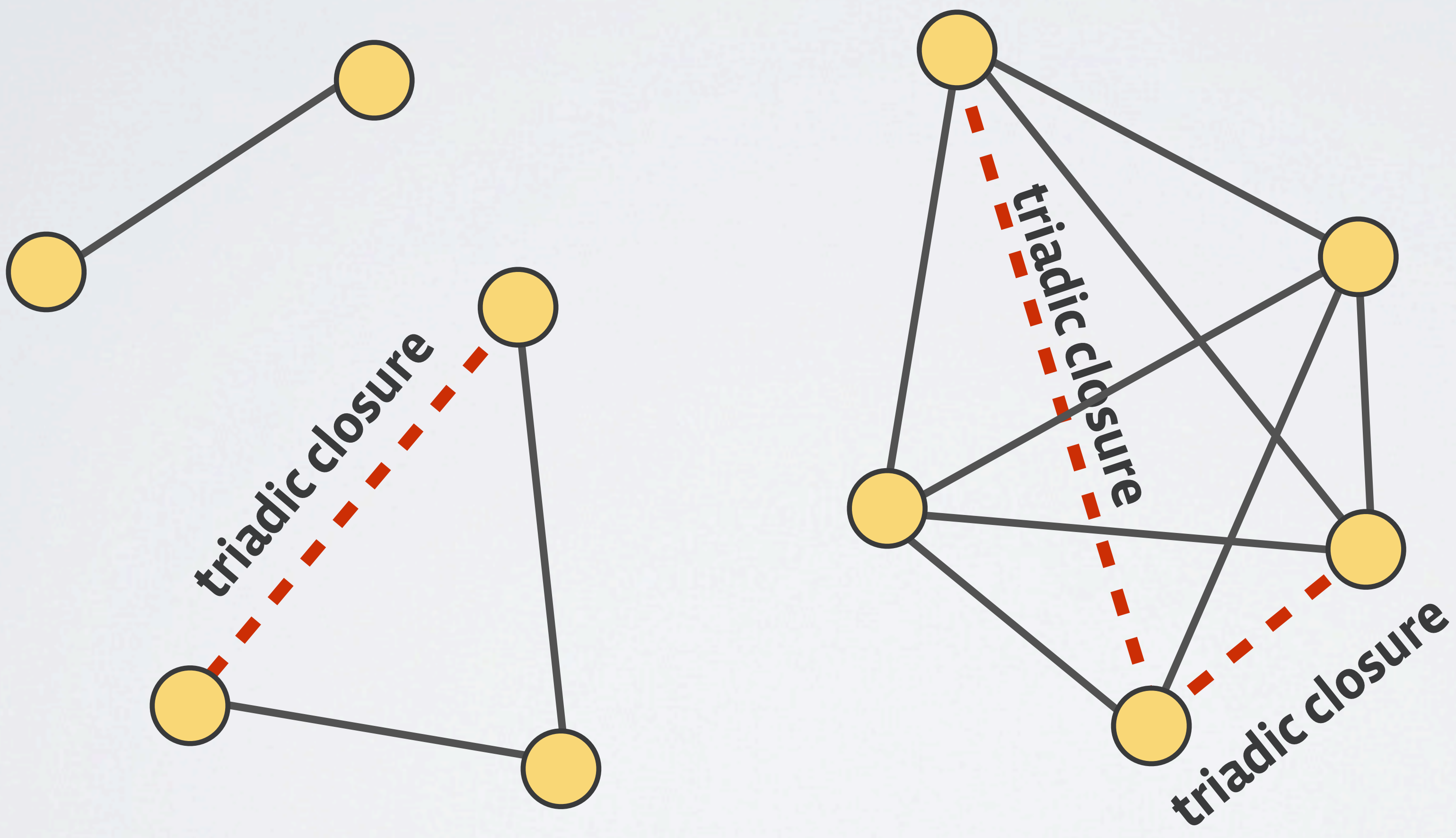
- **Extremal Geography:**
  - How much of this space is even feasible, **combinatorially**?
  - Do empirical graphs fill the **feasible space**?

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

What do we expect?

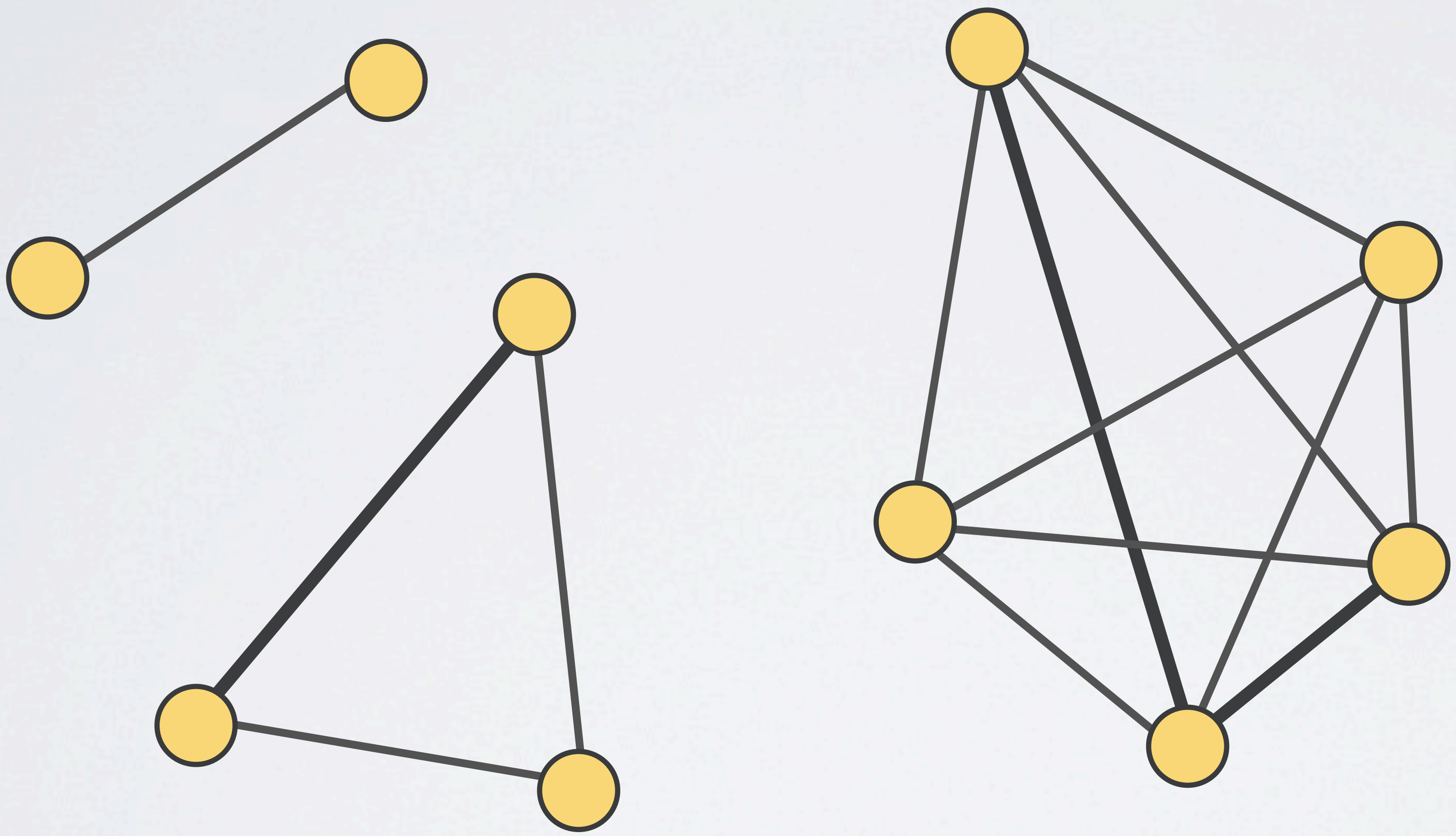


What do we expect?

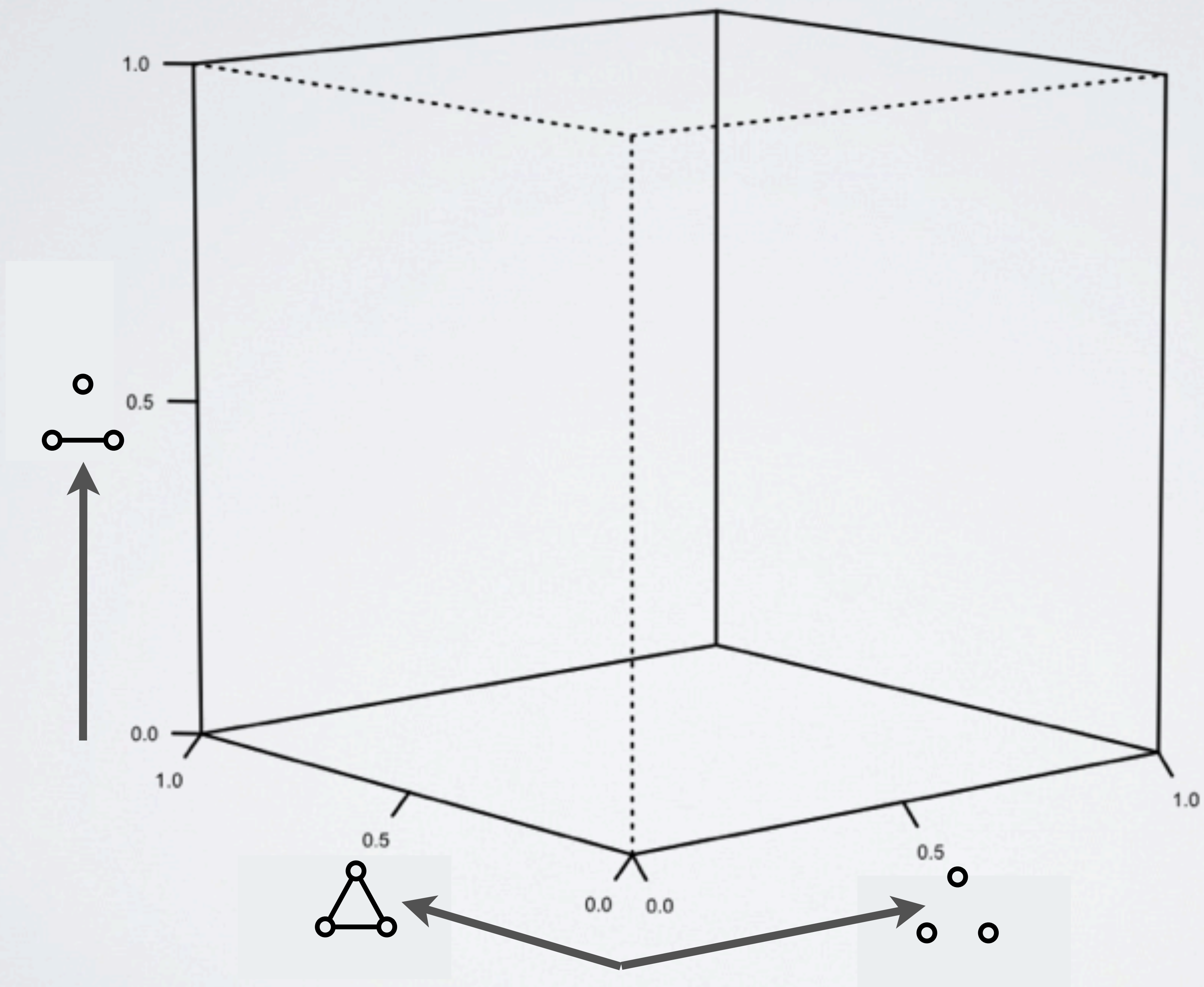


What do we expect?

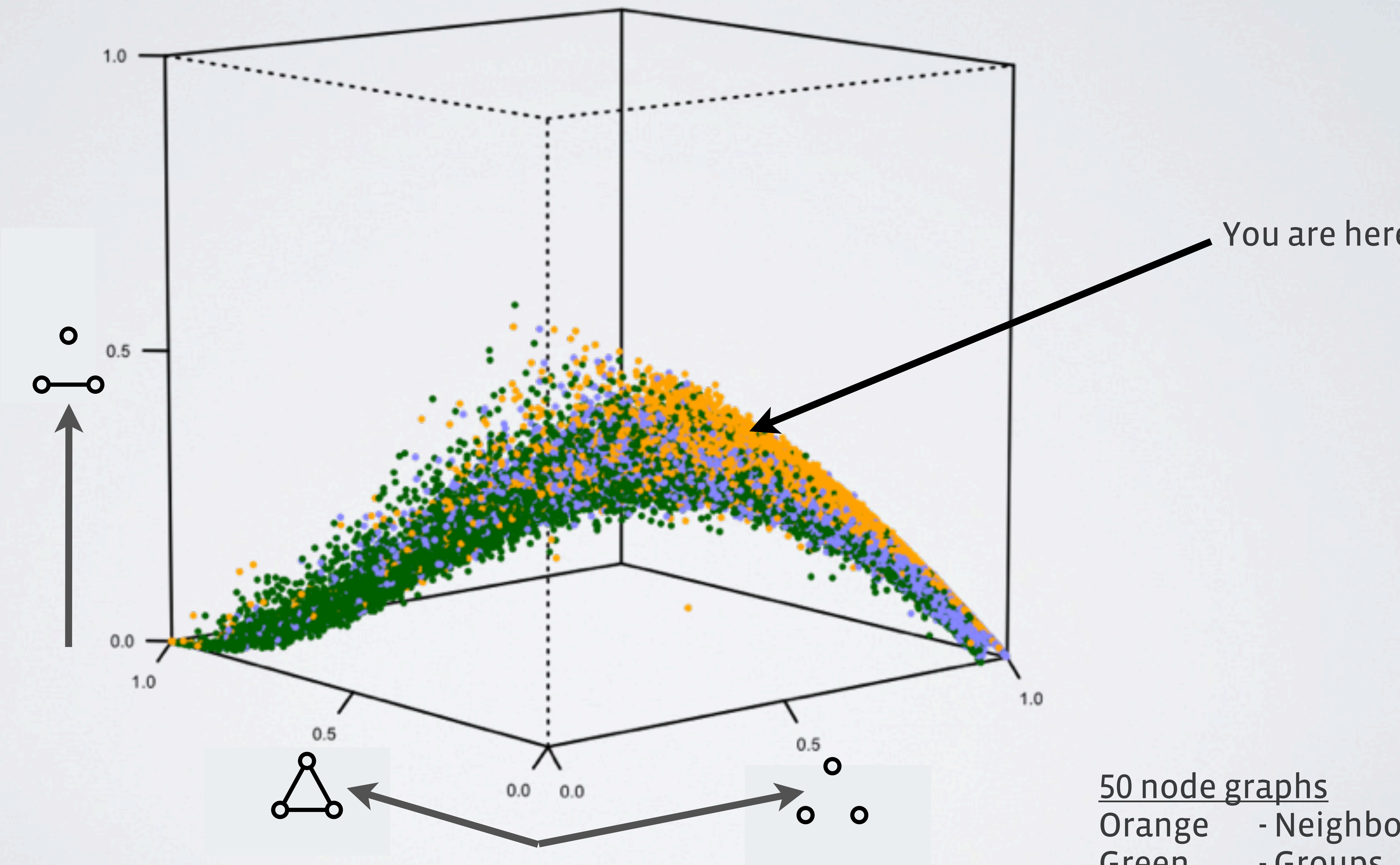
We expect **few wedges, many triangles** for social networks.



The triad space

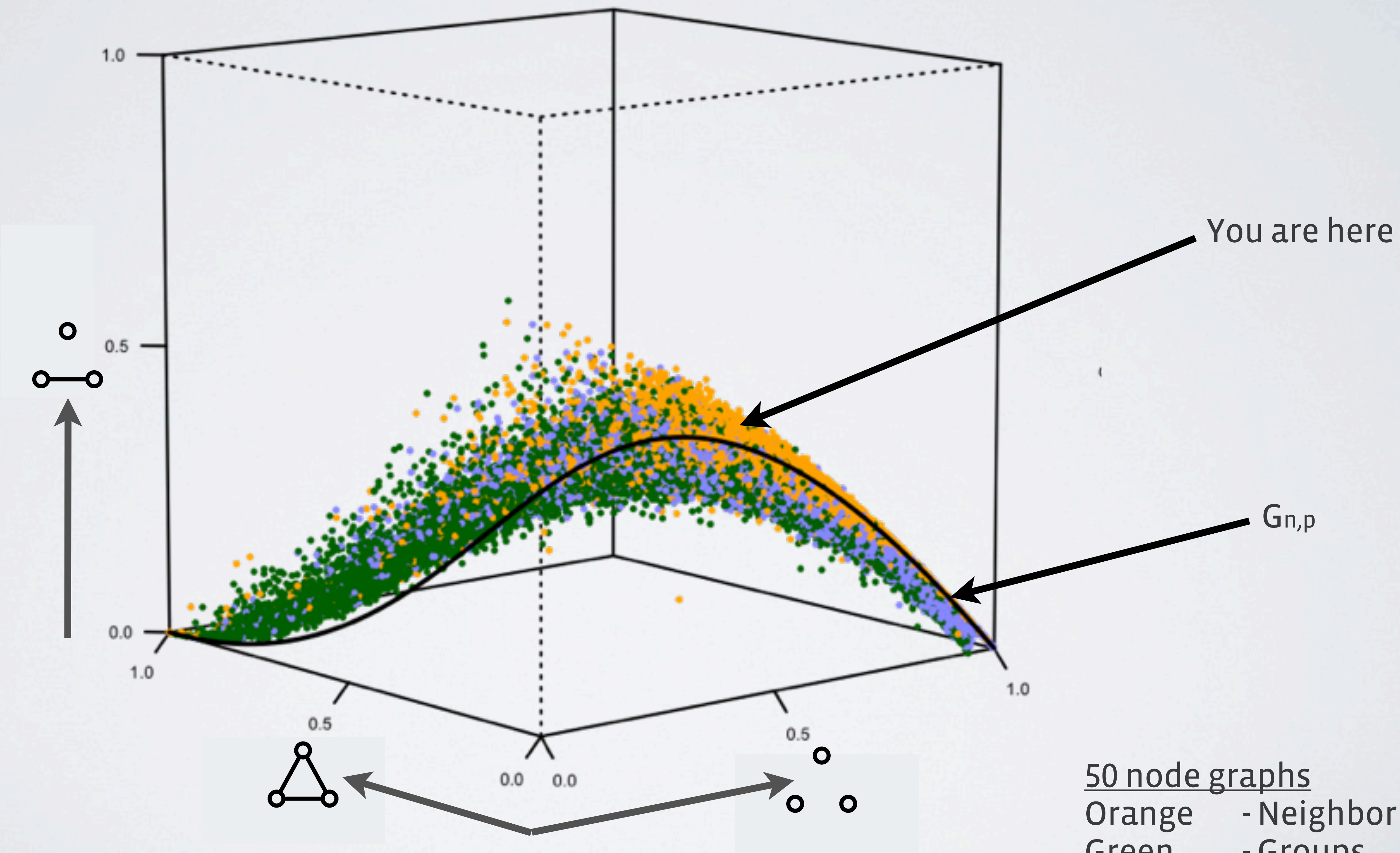


The triad space

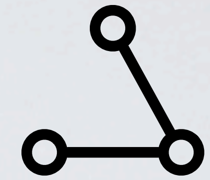


<u>50 node graphs</u>
Orange - Neighborhoods
Green - Groups
Lavender - Events

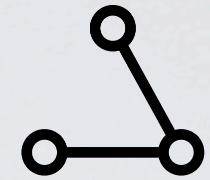
The triad space

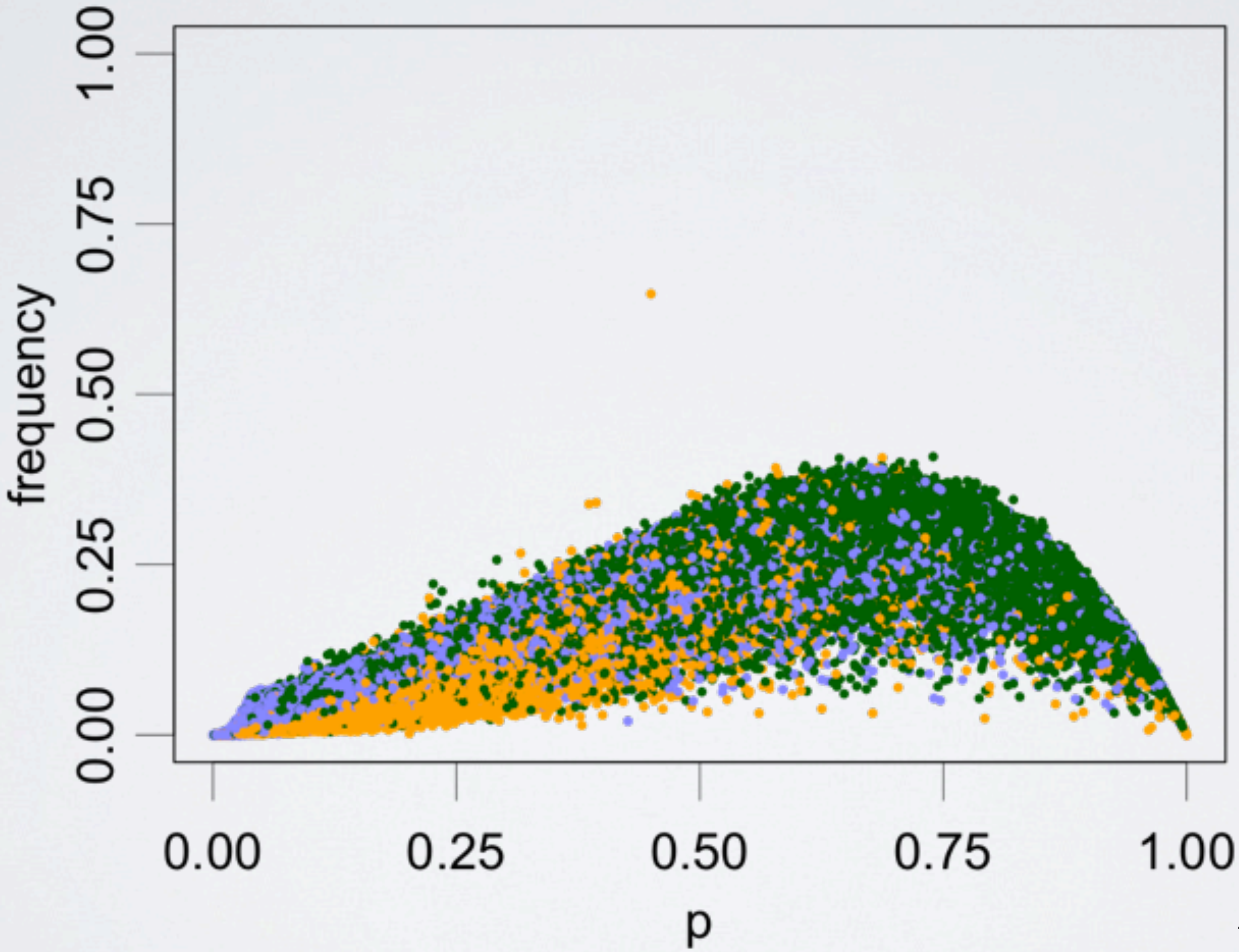


<u>50 node graphs</u>
Orange - Neighborhoods
Green - Groups
Lavender - Events

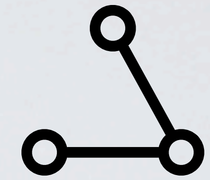
Subgraph frequency of 

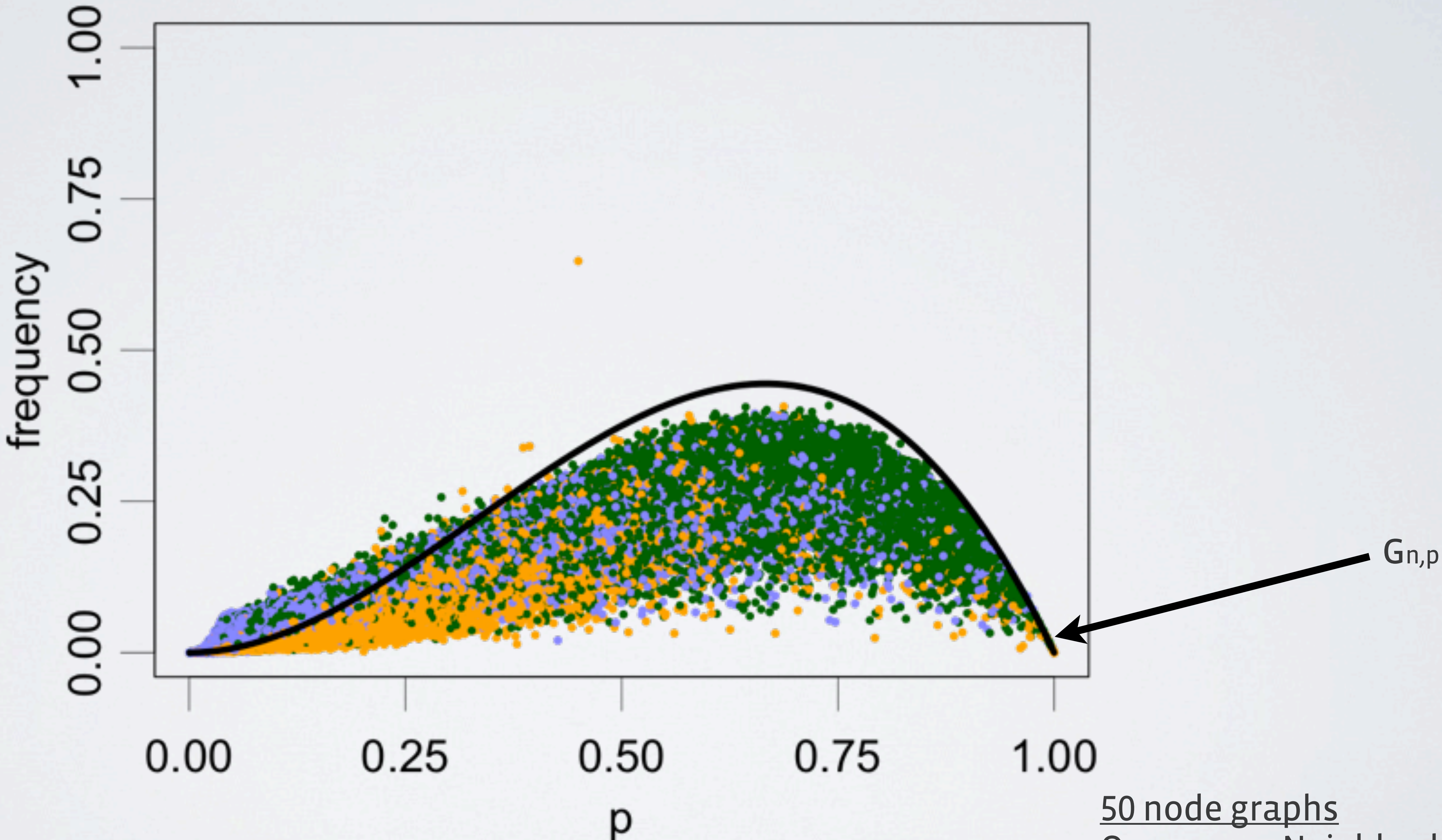
<u>50 node graphs</u>
Orange - Neighborhoods
Green - Groups
Lavender - Events

Subgraph frequency of 

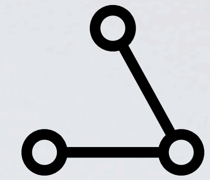


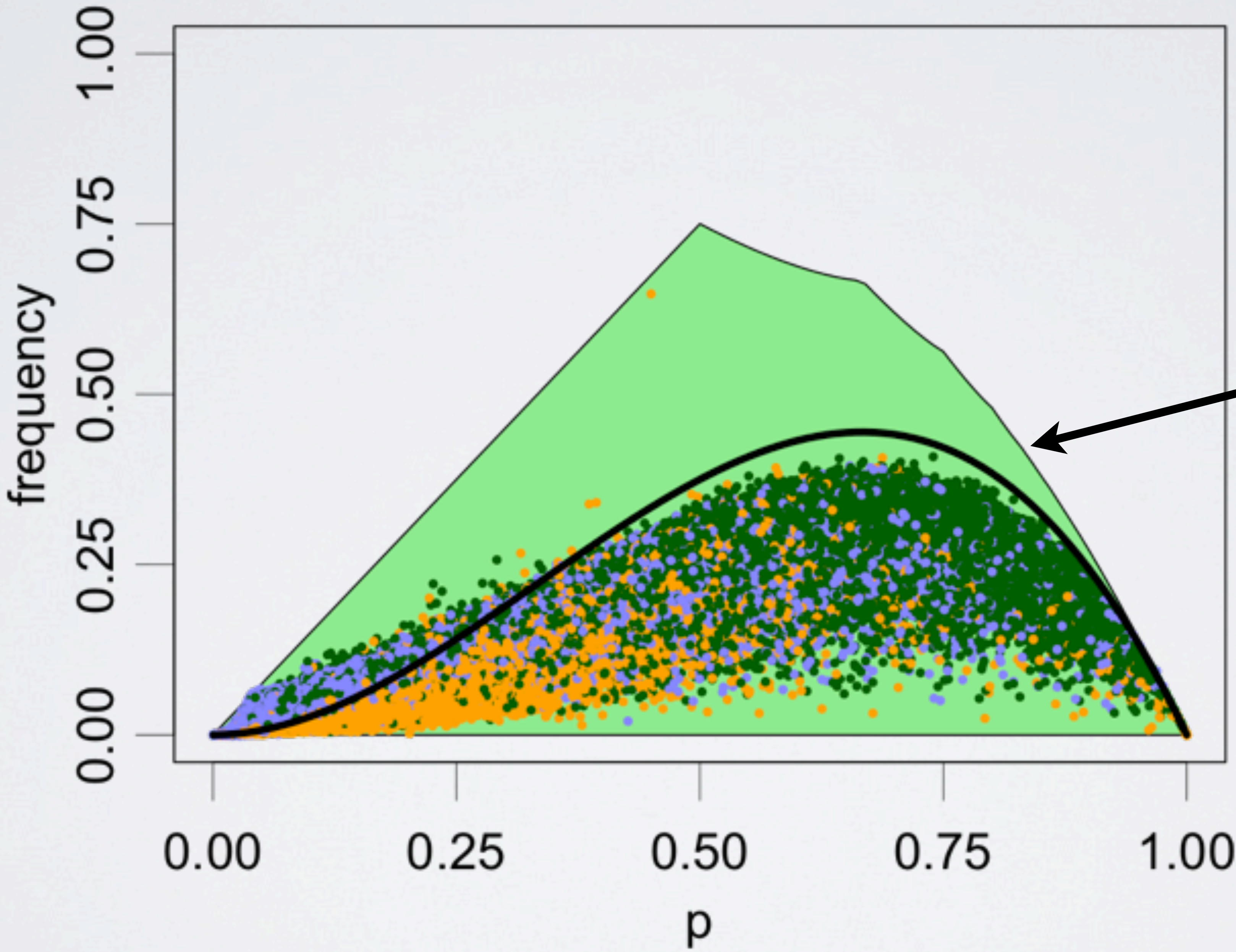
<u>50 node graphs</u>
Orange - Neighborhoods
Green - Groups
Lavender - Events

Subgraph frequency of 



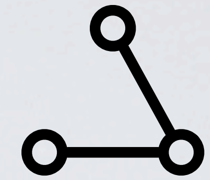
<u>50 node graphs</u>
Orange - Neighborhoods
Green - Groups
Lavender - Events

Subgraph frequency of 

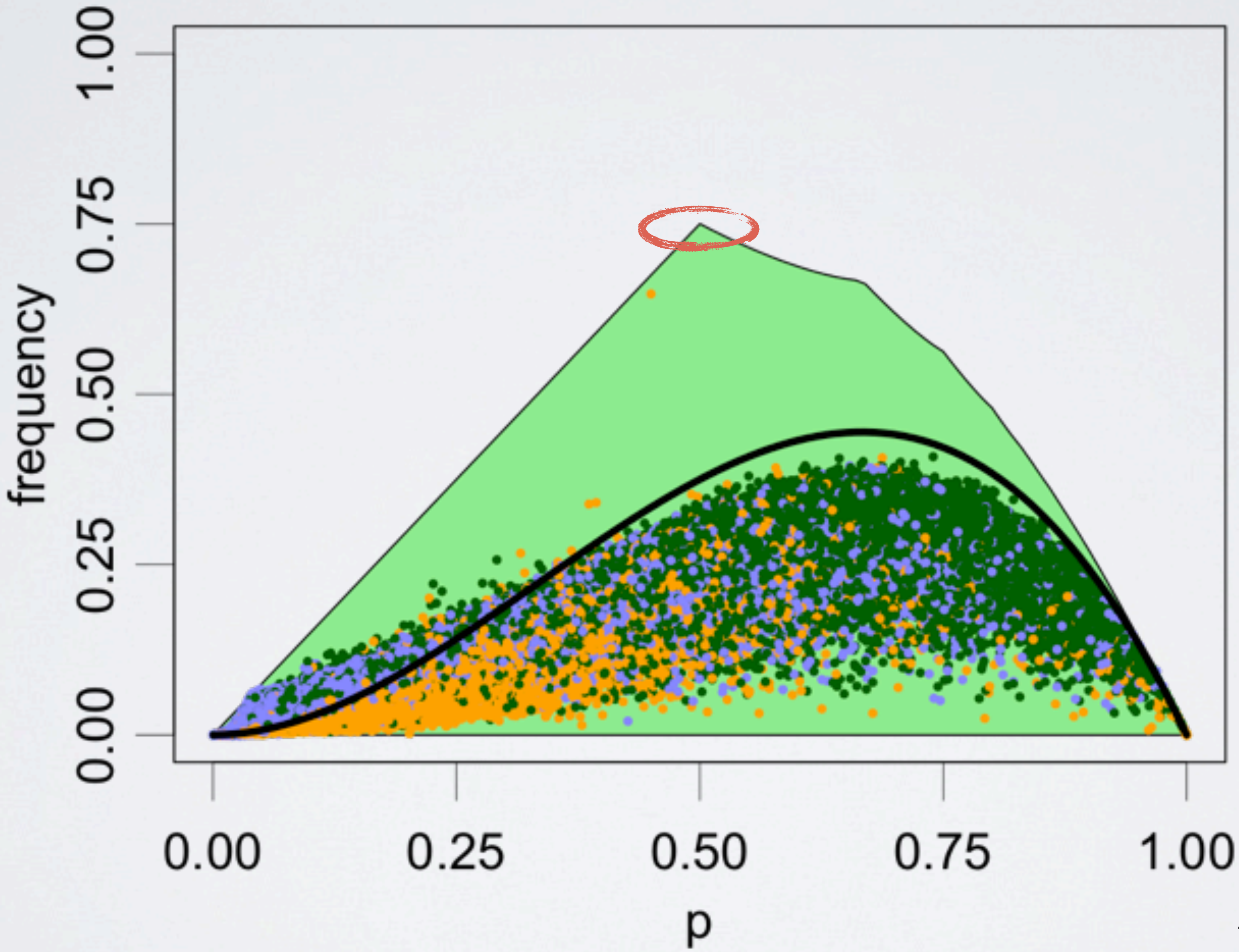


Extremal
Graph Theory

<u>50 node graphs</u>
Orange - Neighborhoods
Green - Groups
Lavender - Events

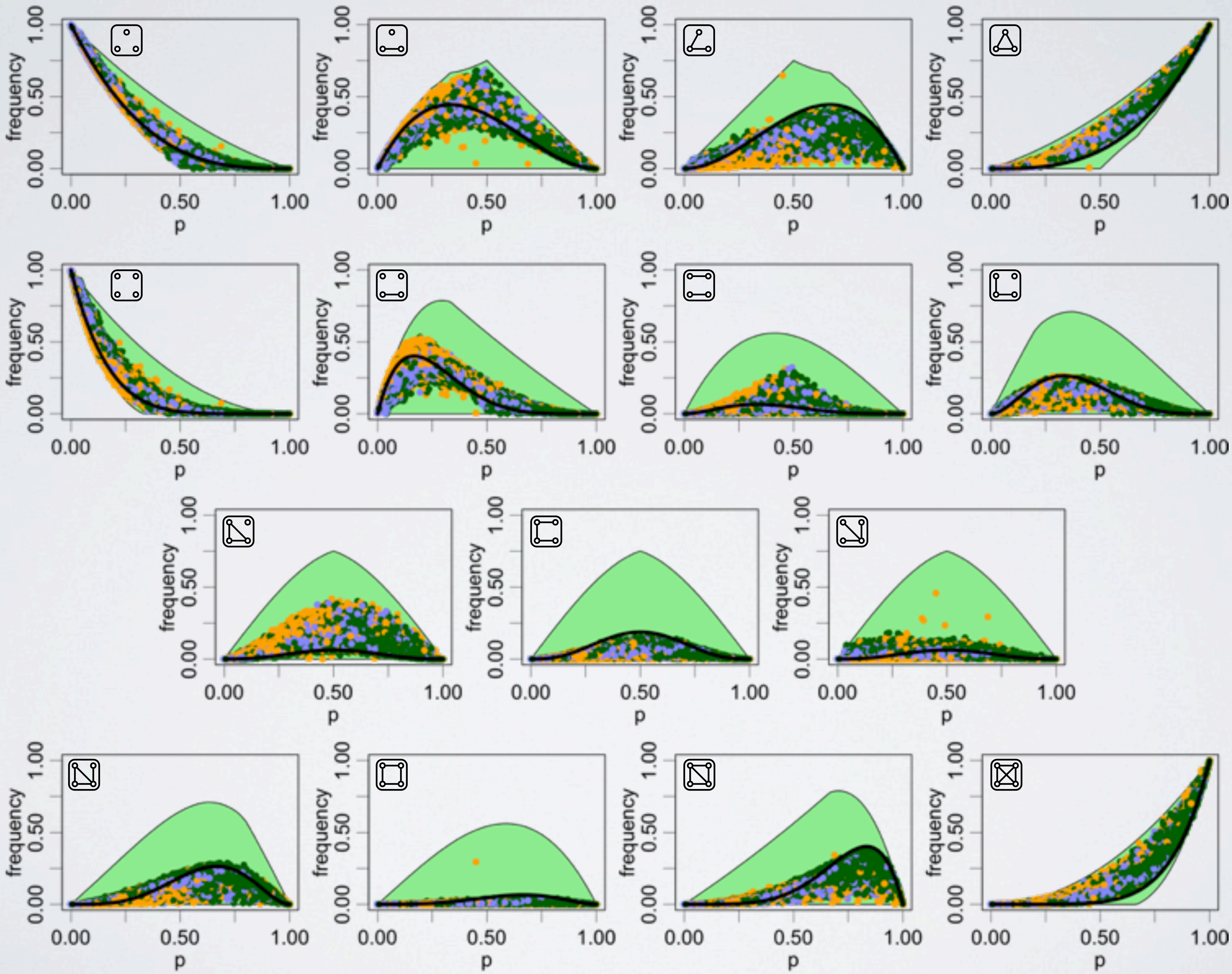
Subgraph frequency of 

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

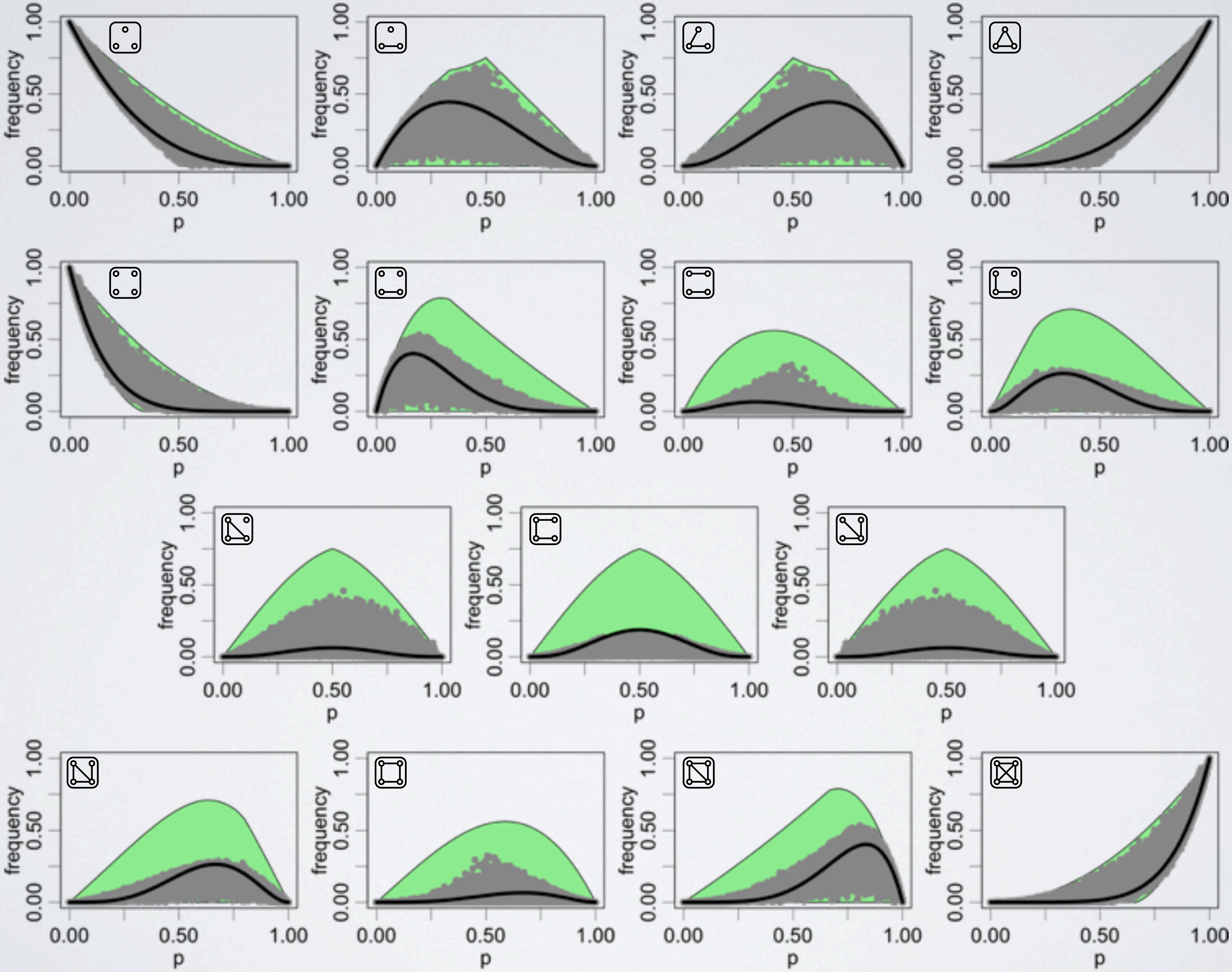


<u>50 node graphs</u>
Orange - Neighborhoods
Green - Groups
Lavender - Events

Subgraph frequencies

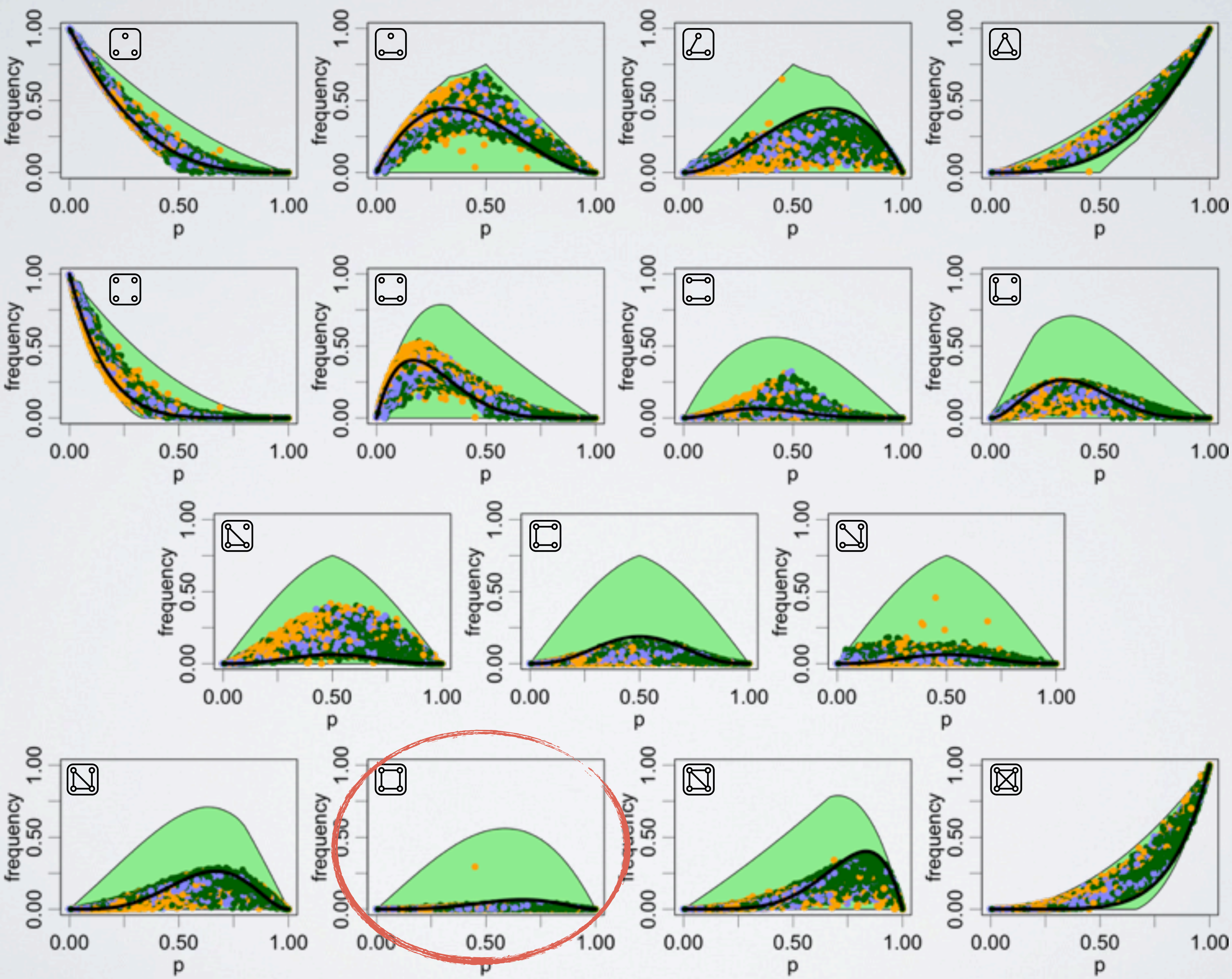


‘Crowd-sourced’ inner bounds



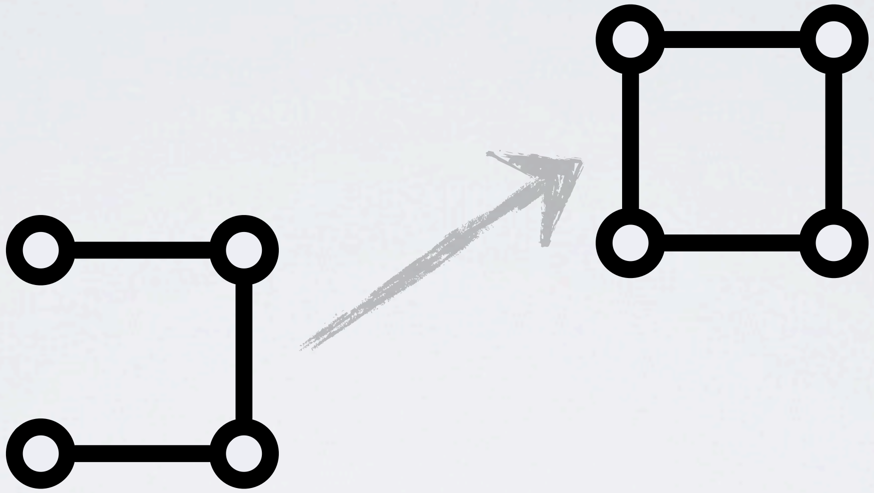
Consider all social graphs and the complements of all graphs, anti-social graphs (which are also graphs!)

What graphs are missing?



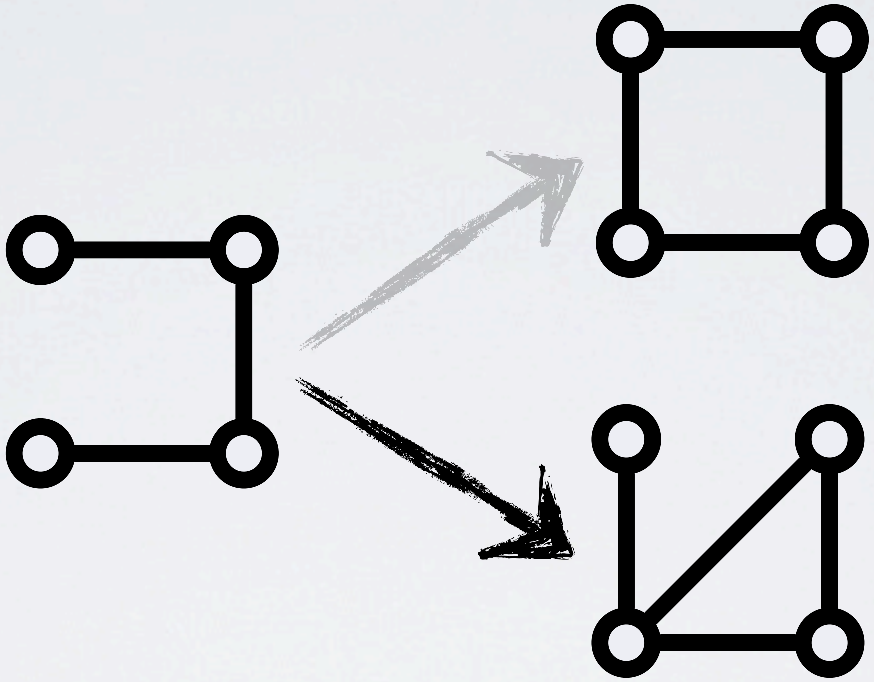
Triadic Closure and Squares

- Square unlikely to form:



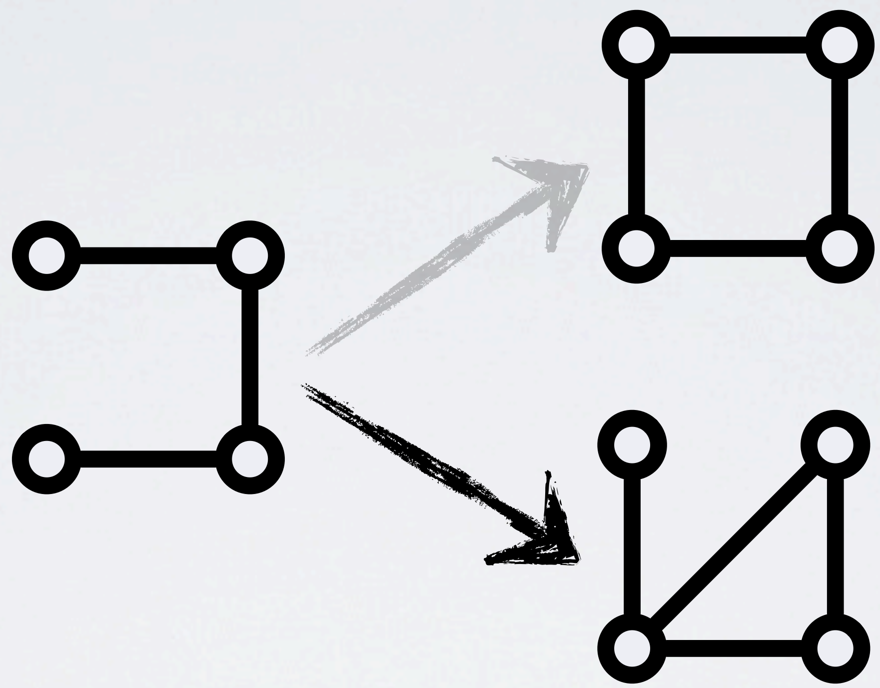
Triadic Closure and Squares

- Square unlikely to form:



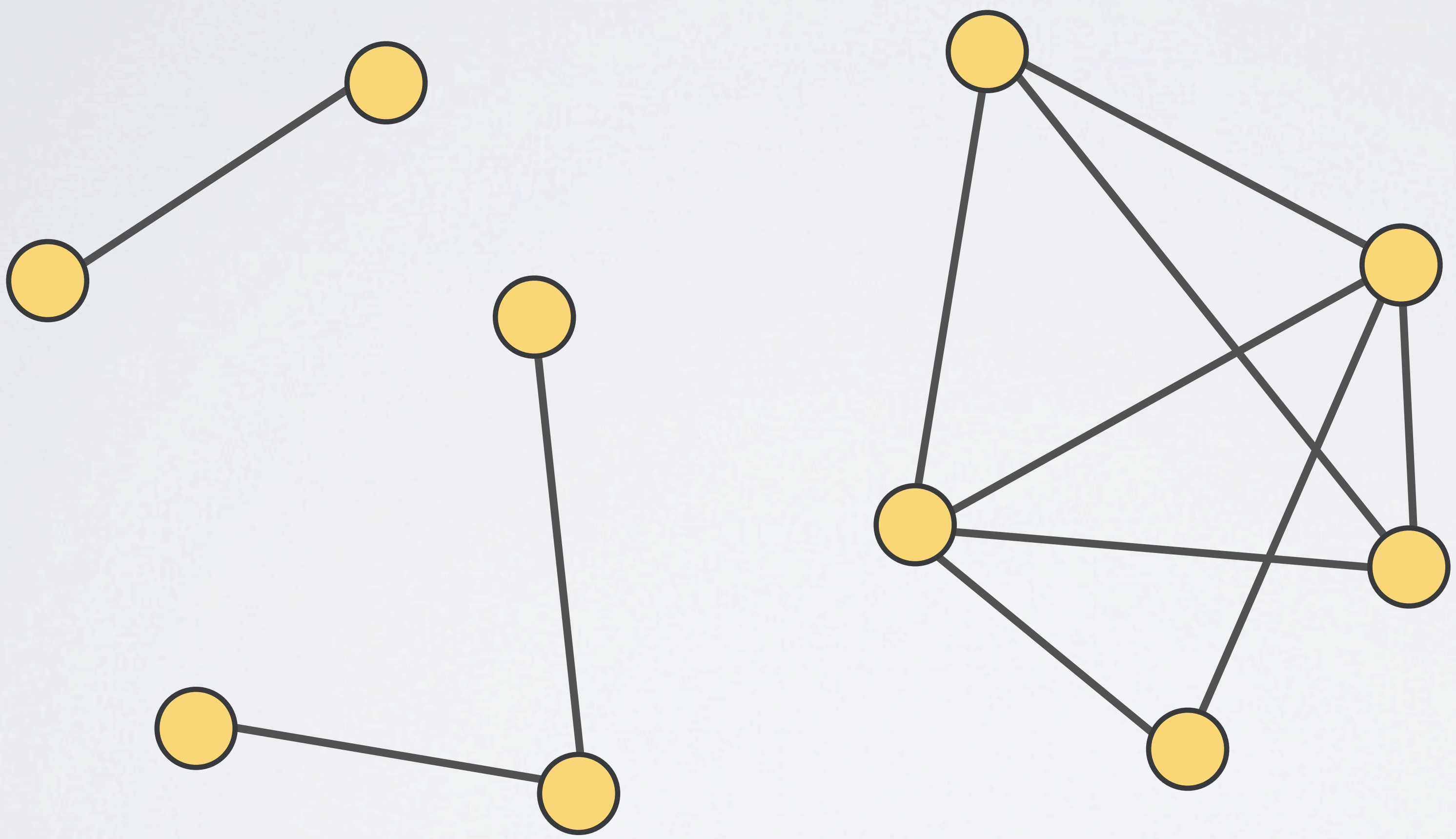
Triadic Closure and Squares

- Square unlikely to form:

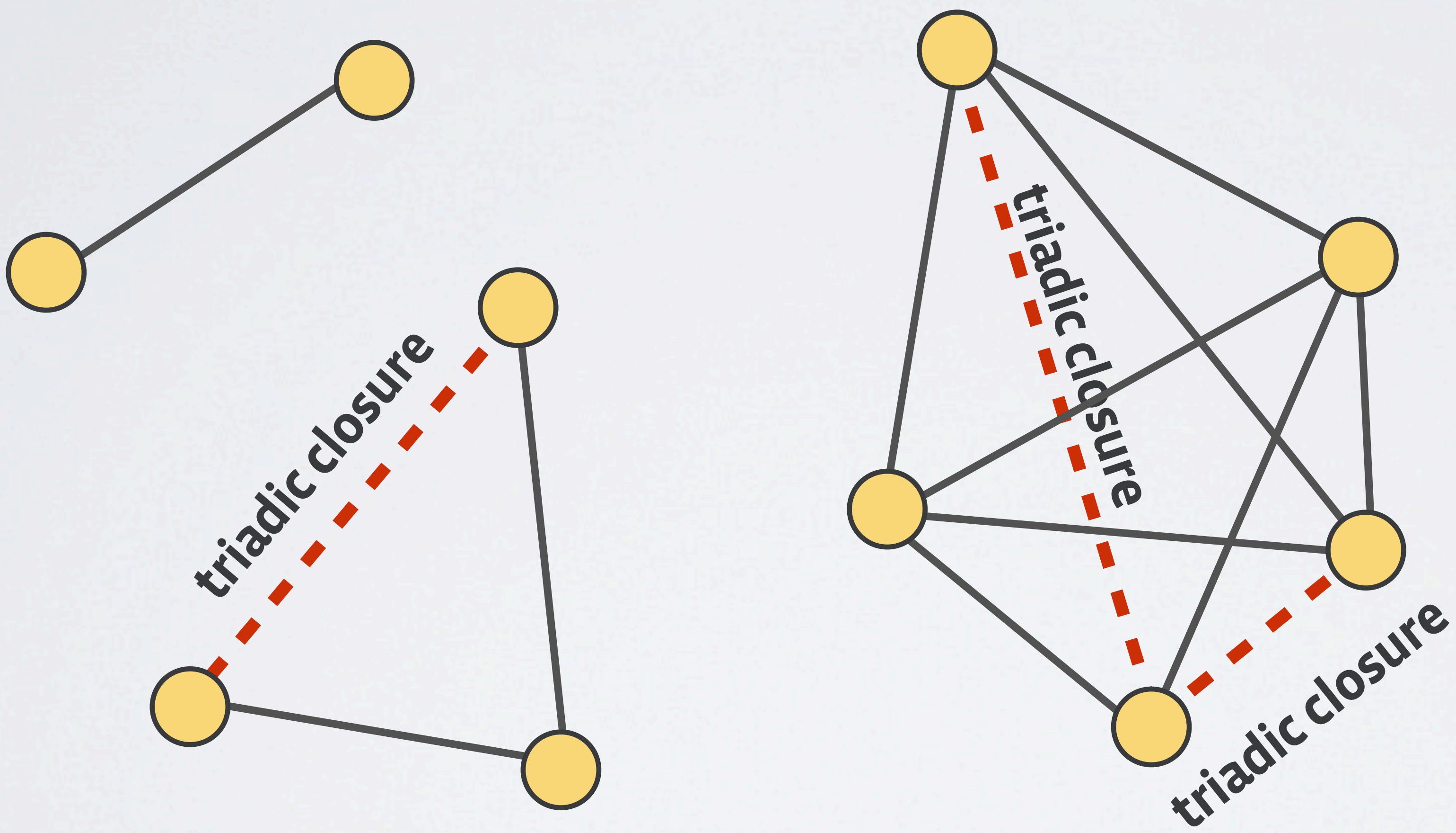


- Square has very short ‘**half-life**’:

Continuous Time Markov Chain Model



Continuous Time Markov Chain Model



Edge Formation Random Walk (EFRW)

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

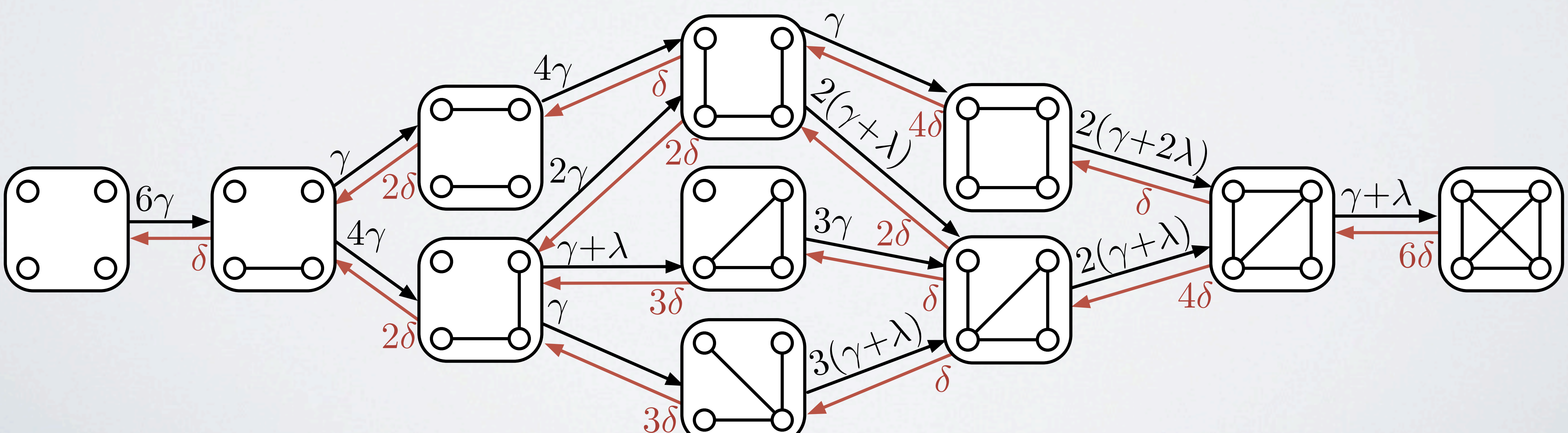
- Independent **Poisson processes** for all node pairs:
  - Arbitrary formation: rate $\gamma > 0$ 
  - Arbitrary deletion: rate $\delta > 0$ 
  - Triadic closure formation for each wedge: rate $\lambda \geq 0$ 

Edge Formation Random Walk (EFRW)

- **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:

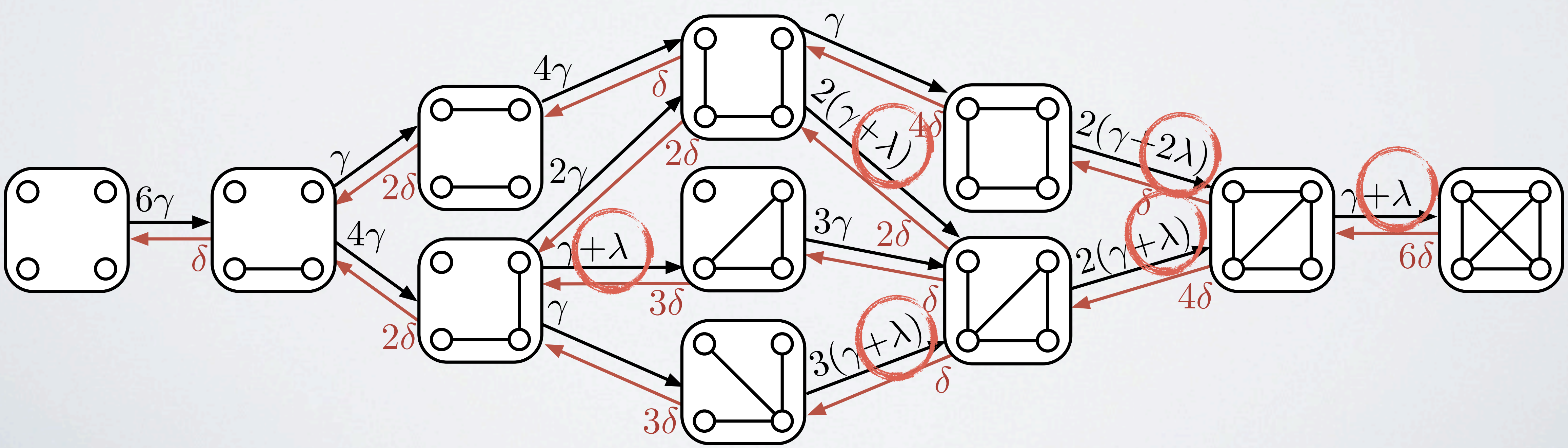


Edge Formation Random Walk (EFRW)

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

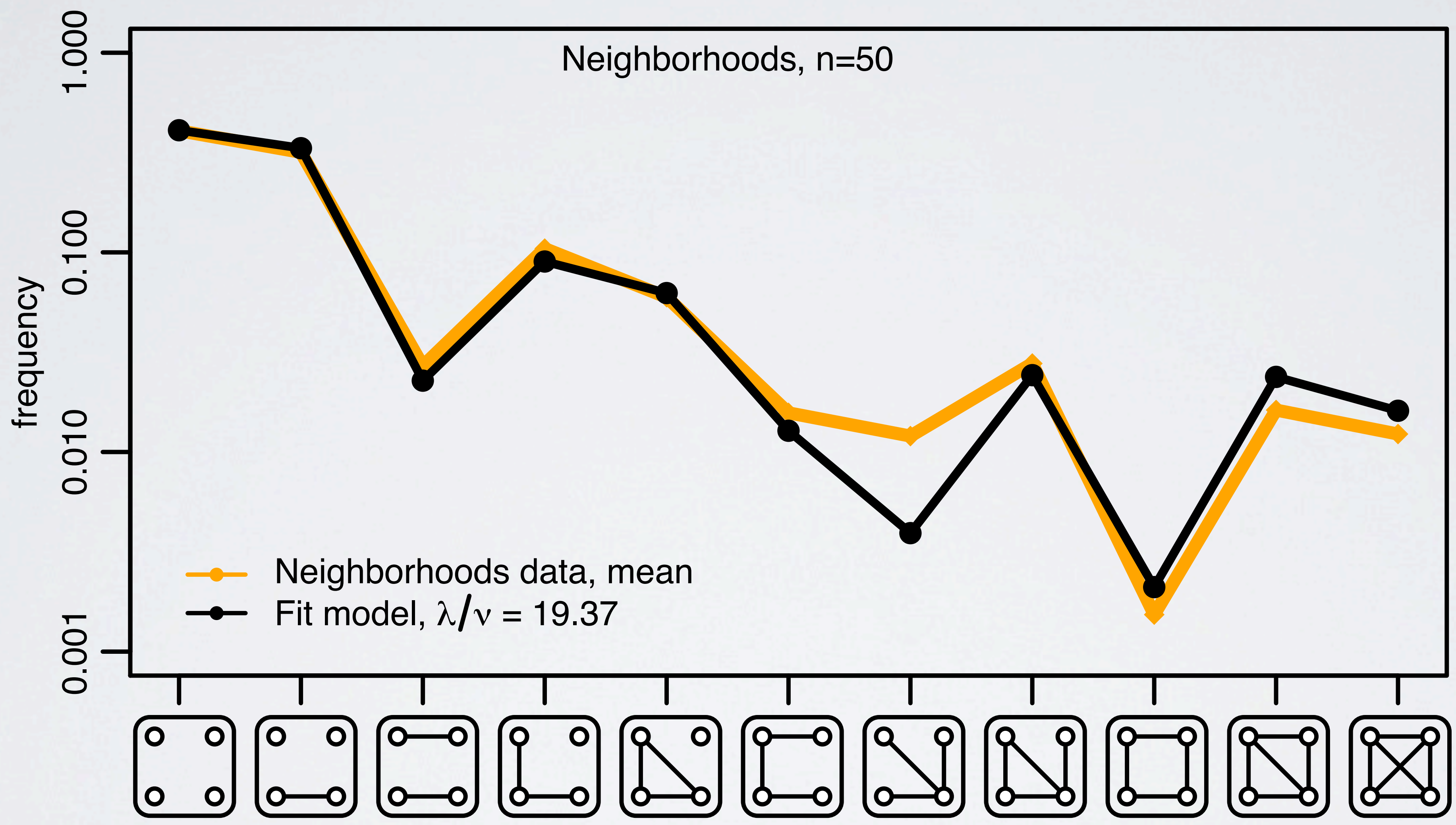
- Independent **Poisson processes** for all node pairs:
  - Arbitrary formation: rate $\gamma > 0$ 
  - Arbitrary deletion: rate $\delta > 0$ 
  - Triadic closure formation for each wedge: rate $\lambda \geq 0$ 

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



Fitting λ to subgraph data

- How well can we fit λ?

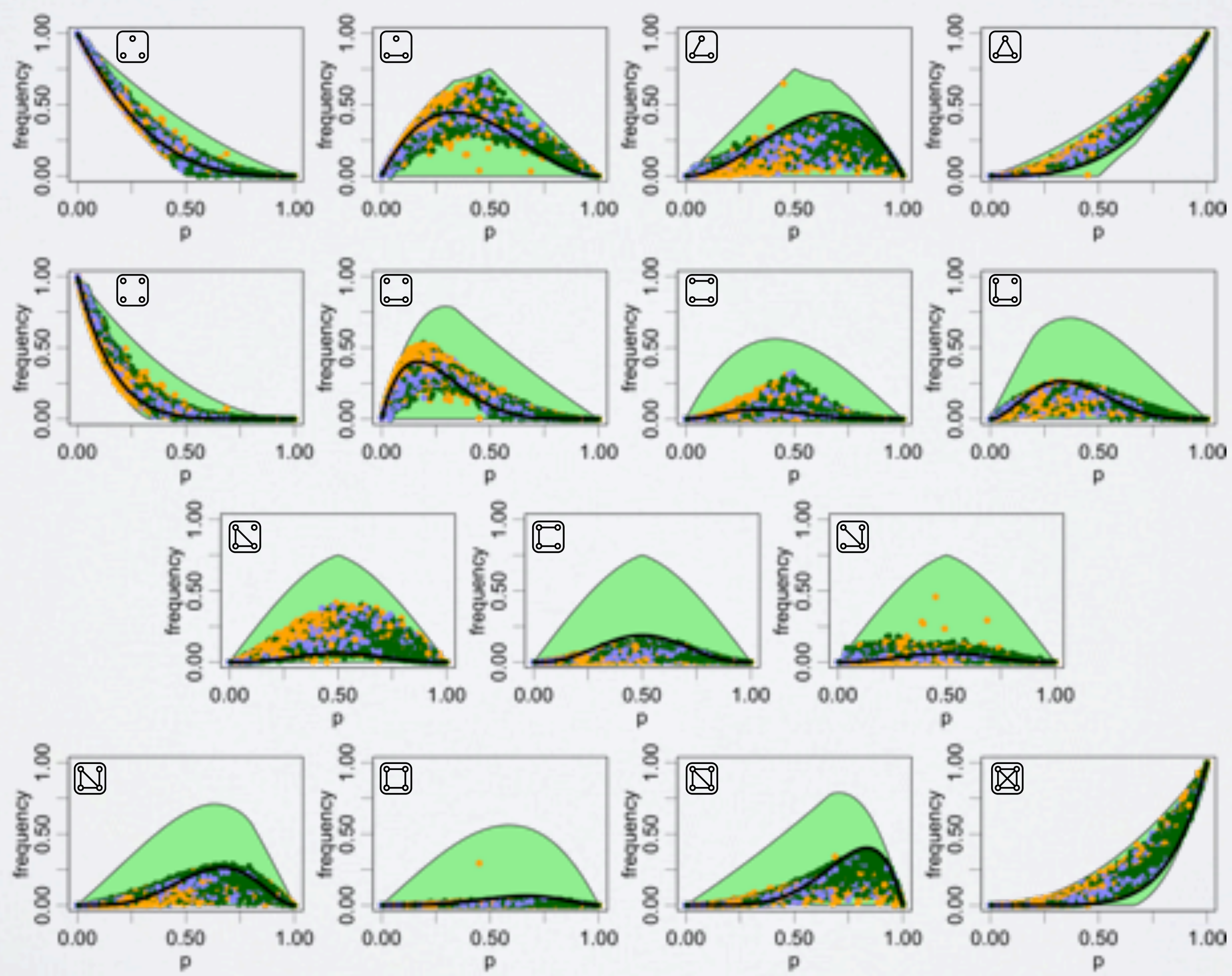


(log-scale y-axis)

- Subgraph frequencies are modeled very well by triadic closure.

Extremal graph theory

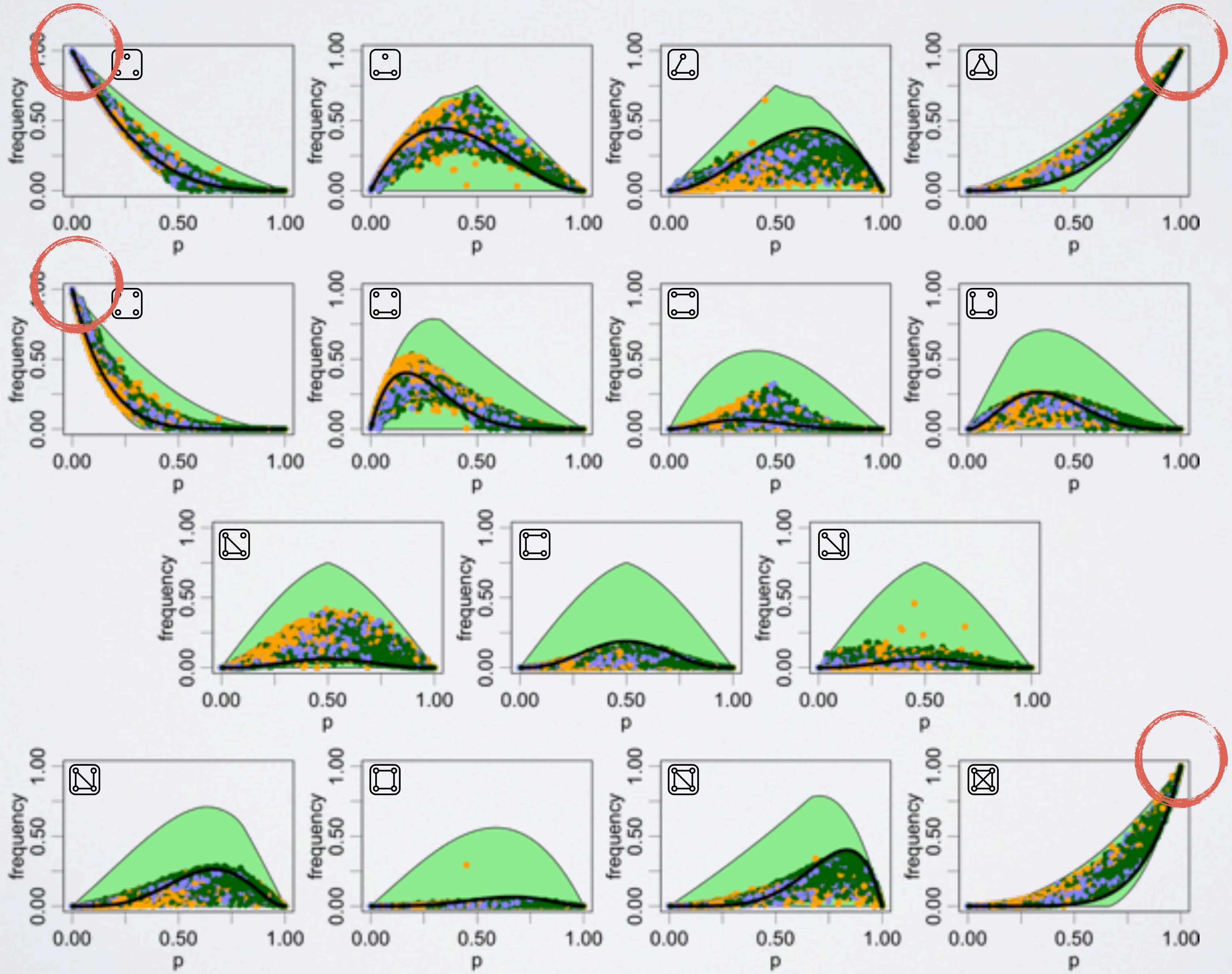
- Subgraph frequencies **s(F,G)** closely related to homomorphism density **t(F,G)**.
  [Borgs et al. 2006, Lovasz 2009]
- Frequency of cliques, lower bounds: **Moon-Moser** 1962, **Razborov** 2008
- Frequency of cliques, upper bounds: **Kruskal-Katona Theorem**
- Frequency of trees: **Sidorenko Conjecture** ('Theorem for trees')
- Also linear relationships across sizes.
- => Linear Program!



Extremal graph theory

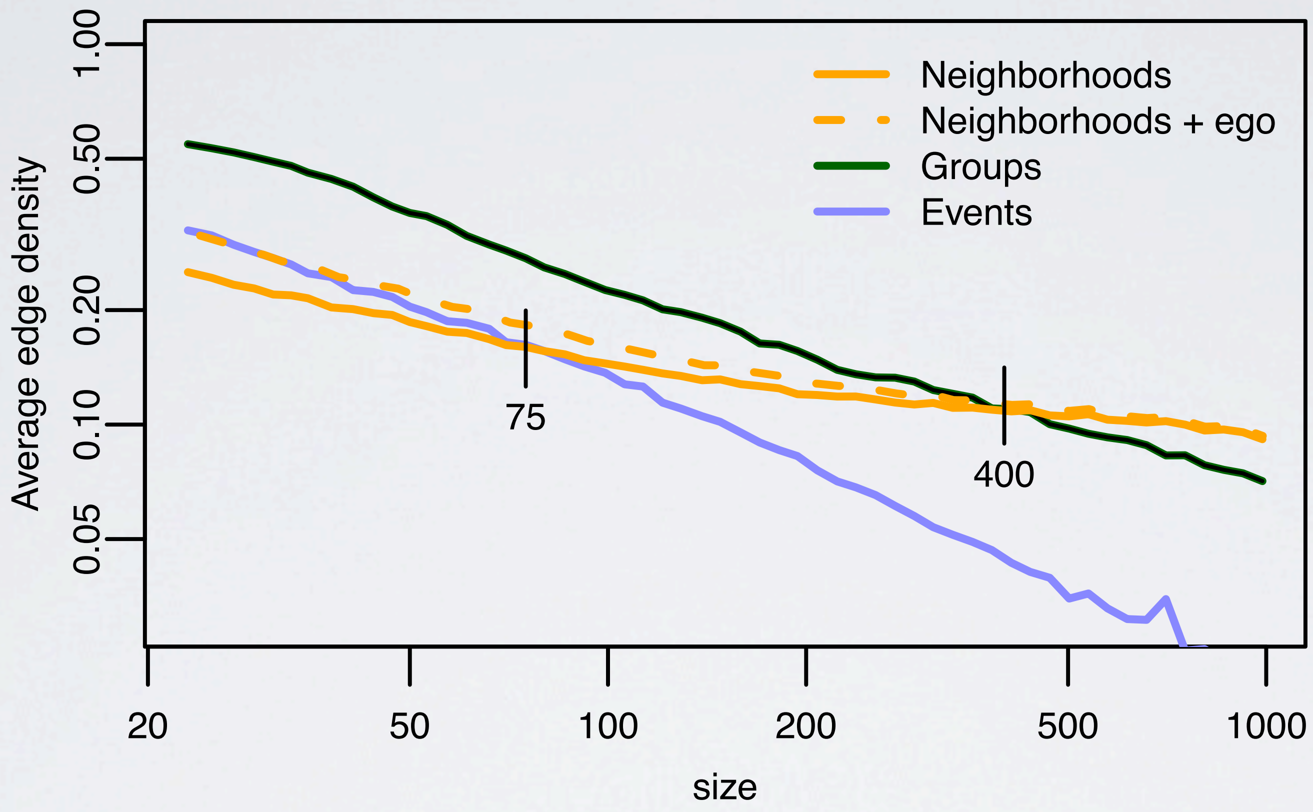
- A proposition for all subgraphs:

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 $s(F, G) < 1 - \epsilon$.



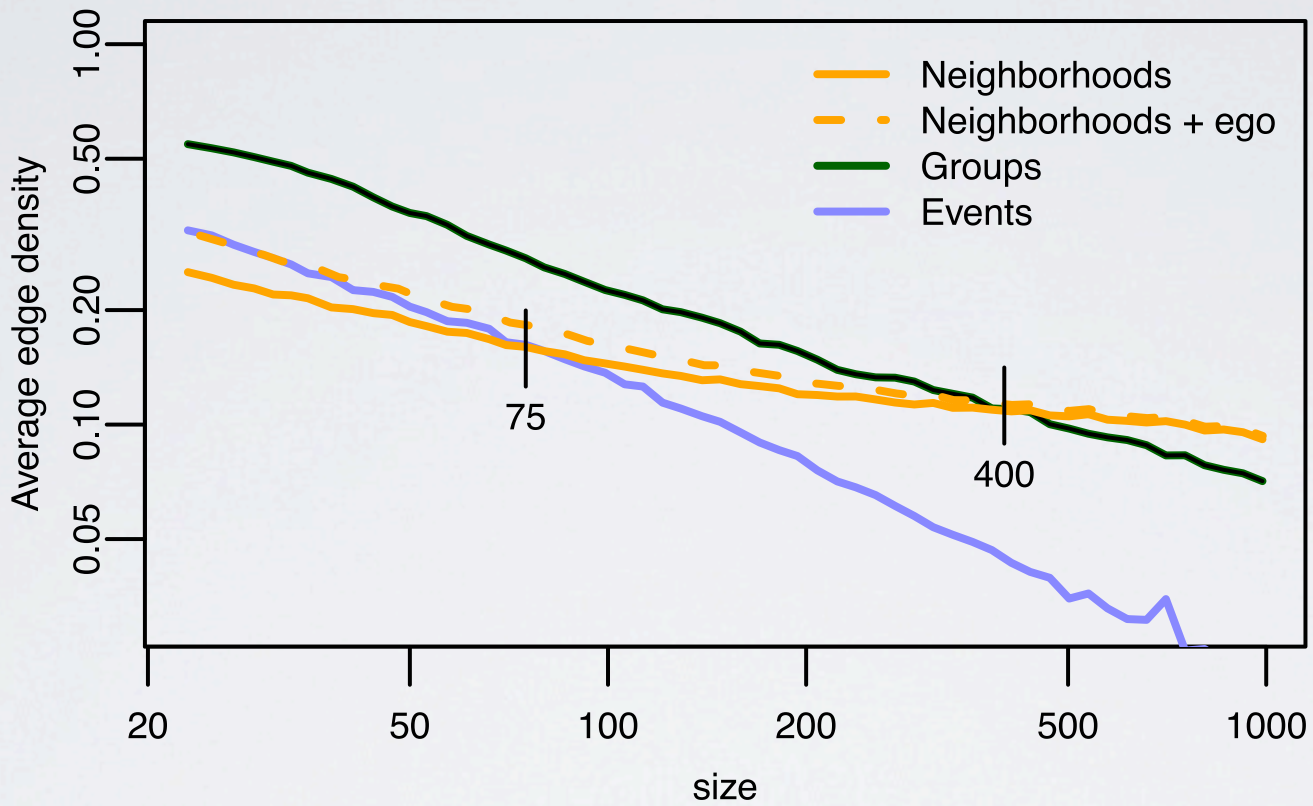
Audience graph classification

- How do different audience graphs differ?



Audience graph classification

- How do different audience graphs differ?

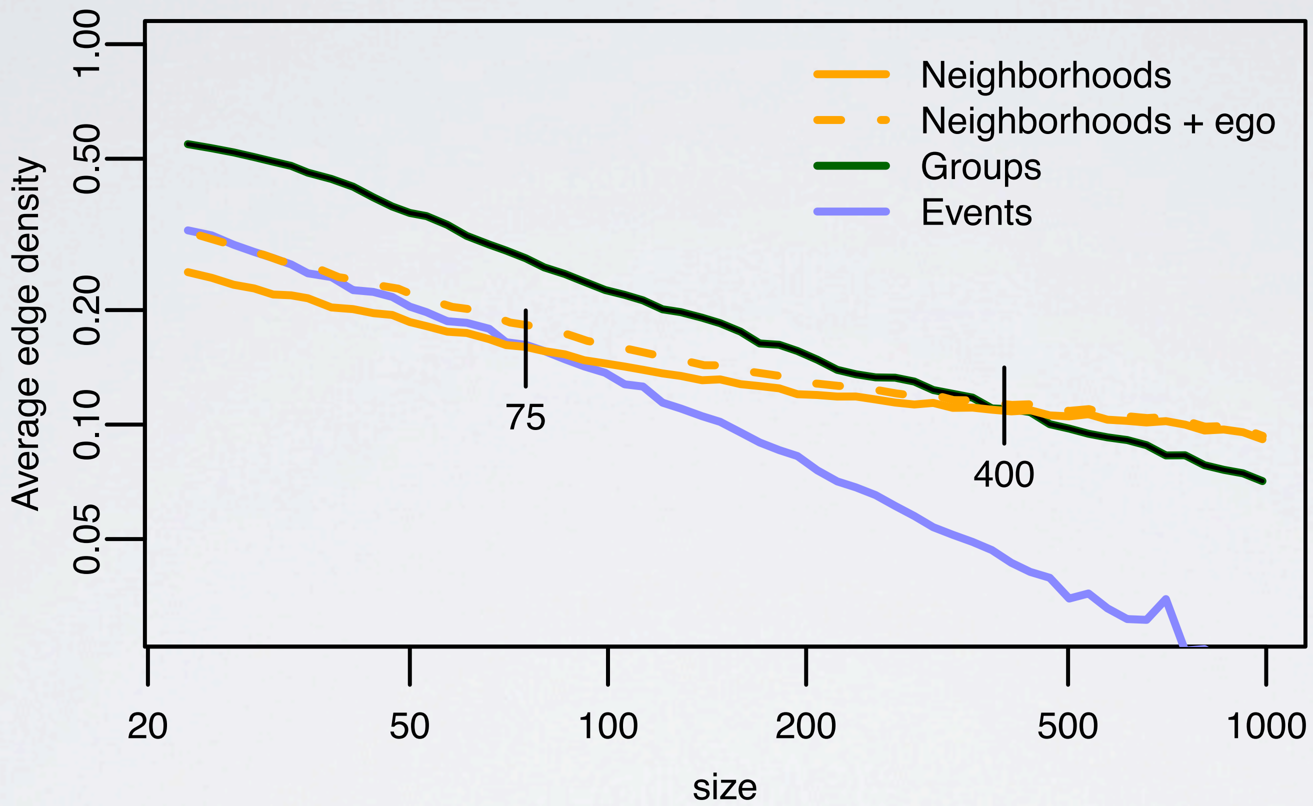


- Classification challenges

A) 75-node neigh. vs. 75-node events
B) 400-node neigh. vs. 400-node groups

Audience graph classification

- How do different audience graphs differ?



- Classification challenges

  A) 75-node neigh. vs. 75-node events

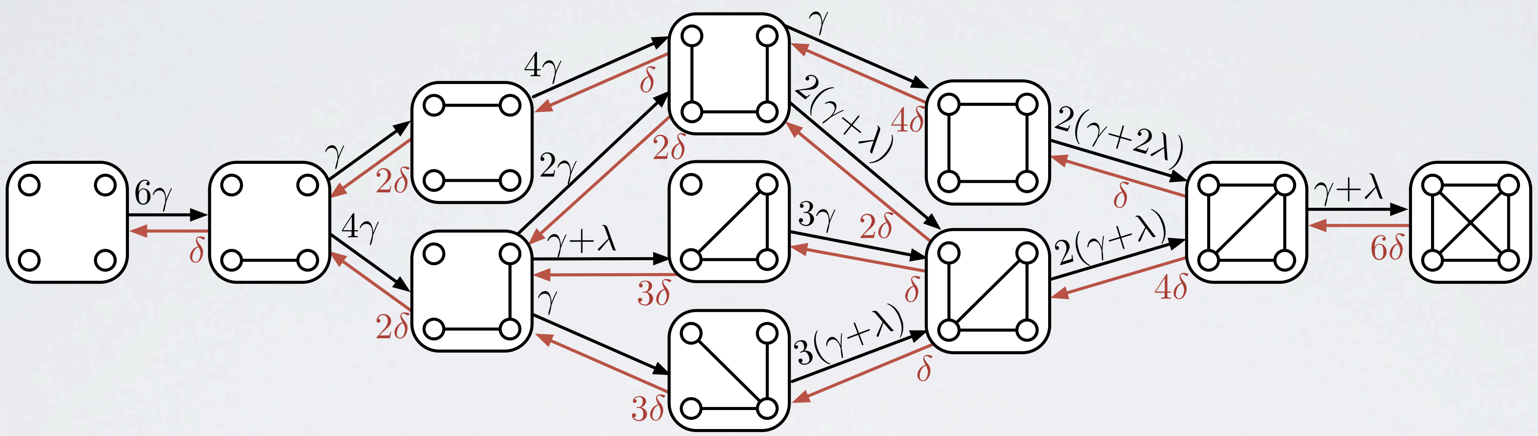
  B) 400-node neigh. vs. 400-node groups

- Features:

Quad frequencies :	76% / 76%	accuracy
Global features:	69% / 76%	accuracy
Quad frequencies + Global features:	81% / 82%	accuracy

Conclusions

- Subgraph frequencies usefully characterize social graphs, have extremal limits!
- Edge Formation Random Walk model of dense social graphs:



- Homomorphism density bounds yield subgraph density bounds: