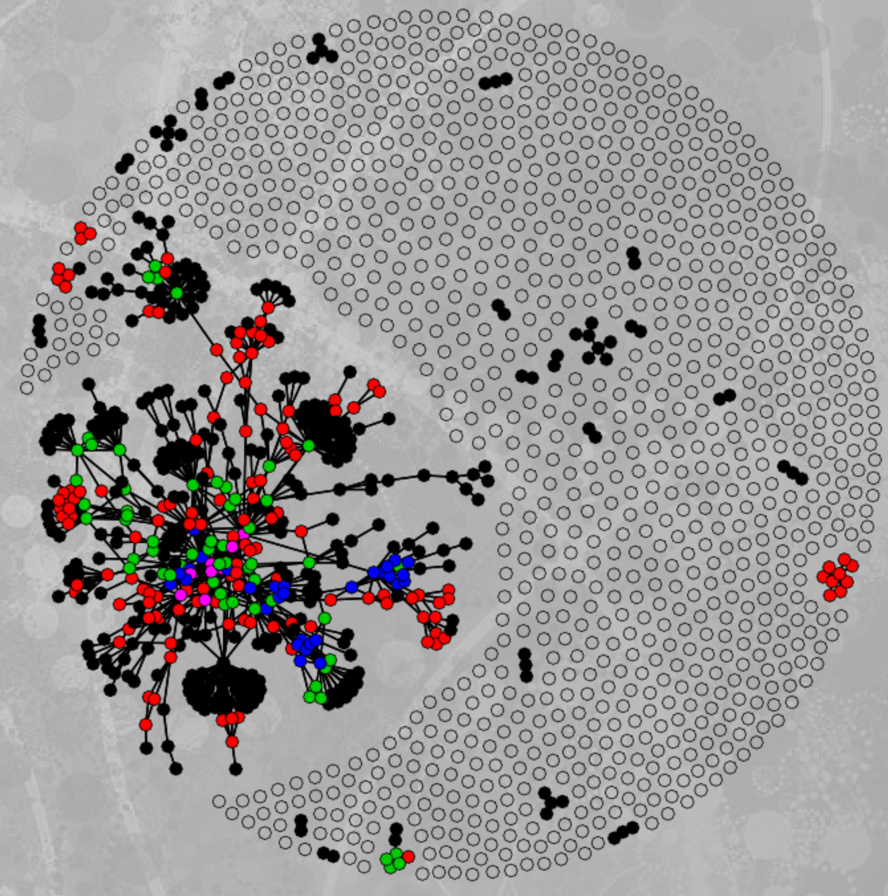
Statistical Models for Network Data: What and Why

Carter T. Butts
Department of Sociology and Institute for Mathematical Behavioral Sciences
University of California, Irvine

Prepared for the December 8, 2009 UCI MURI AHM. This work was supported by DOD ONR award N00014-8-1-1015.

Introduction: The "What"

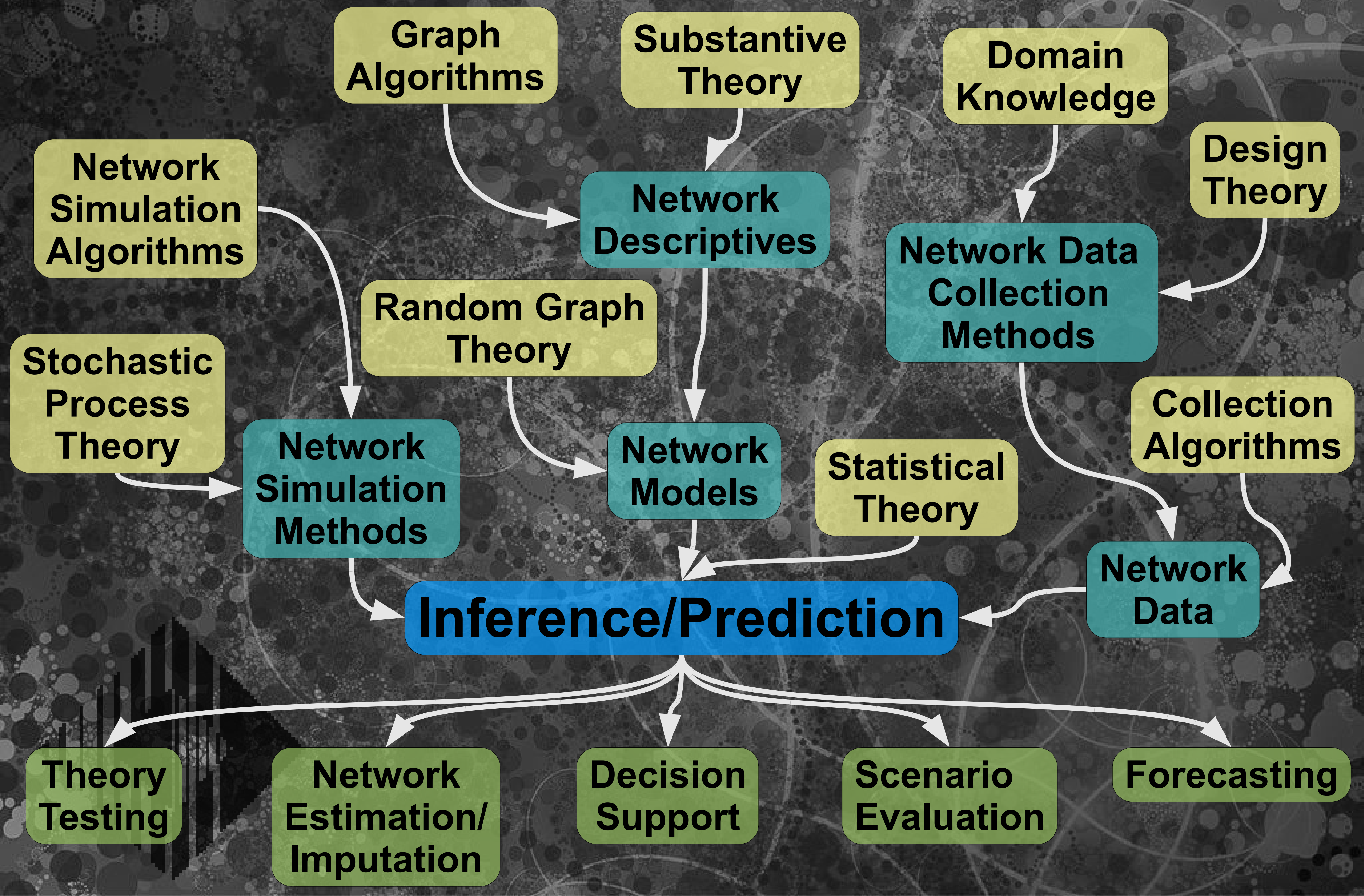
- The key questions regarding sociotechnical systems are relational
  - Connectivity, robustness, centrality, diffusion, etc.
- How do we make sense of this information?
- The statistical approach:
  - Assume that what we see reflects processes with many potential outcomes
  - Posit models that reflect our uncertainty about unknowns
  - Reason from observations and prior knowledge to unknown quantities in a principled manner



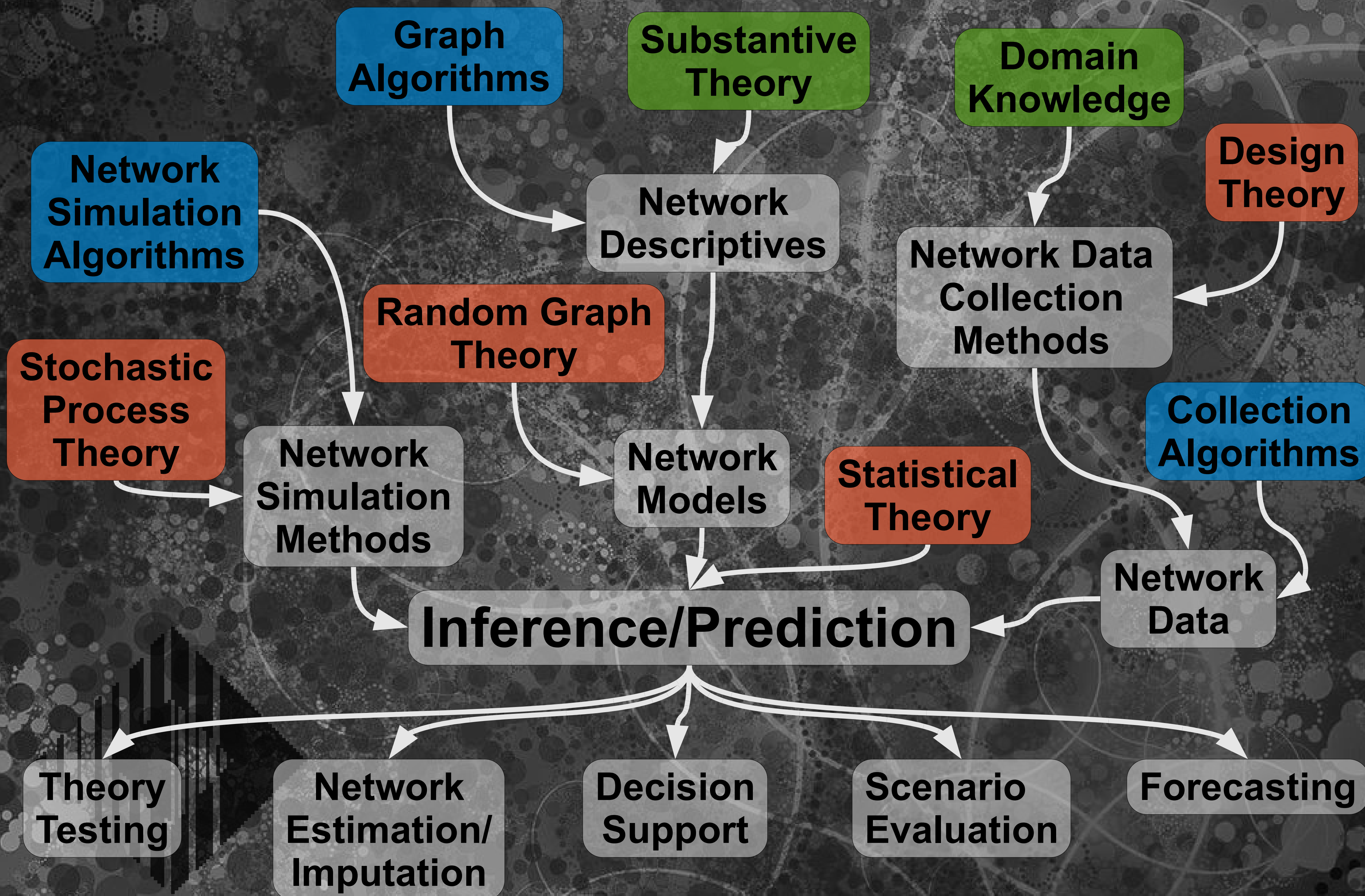
Key Challenges for this Approach

- Parameterizing models in a sensible and computable way
  - Models must reflect phenomenological understanding, but must also scale to real data
- Accounting for data collection
  - Need sampling methods, ways of handling missing/error-prone data
- Making inference both principled *and* practical
  - Want accurate estimates, but can't wait forever for results
- Dealing with rich, dynamic data
  - Real-world problems involve systems with complex covariates (text, geography, etc.) that change over time
- ***In sum: statistically principled methods that respect the realities of data and computational constraints***

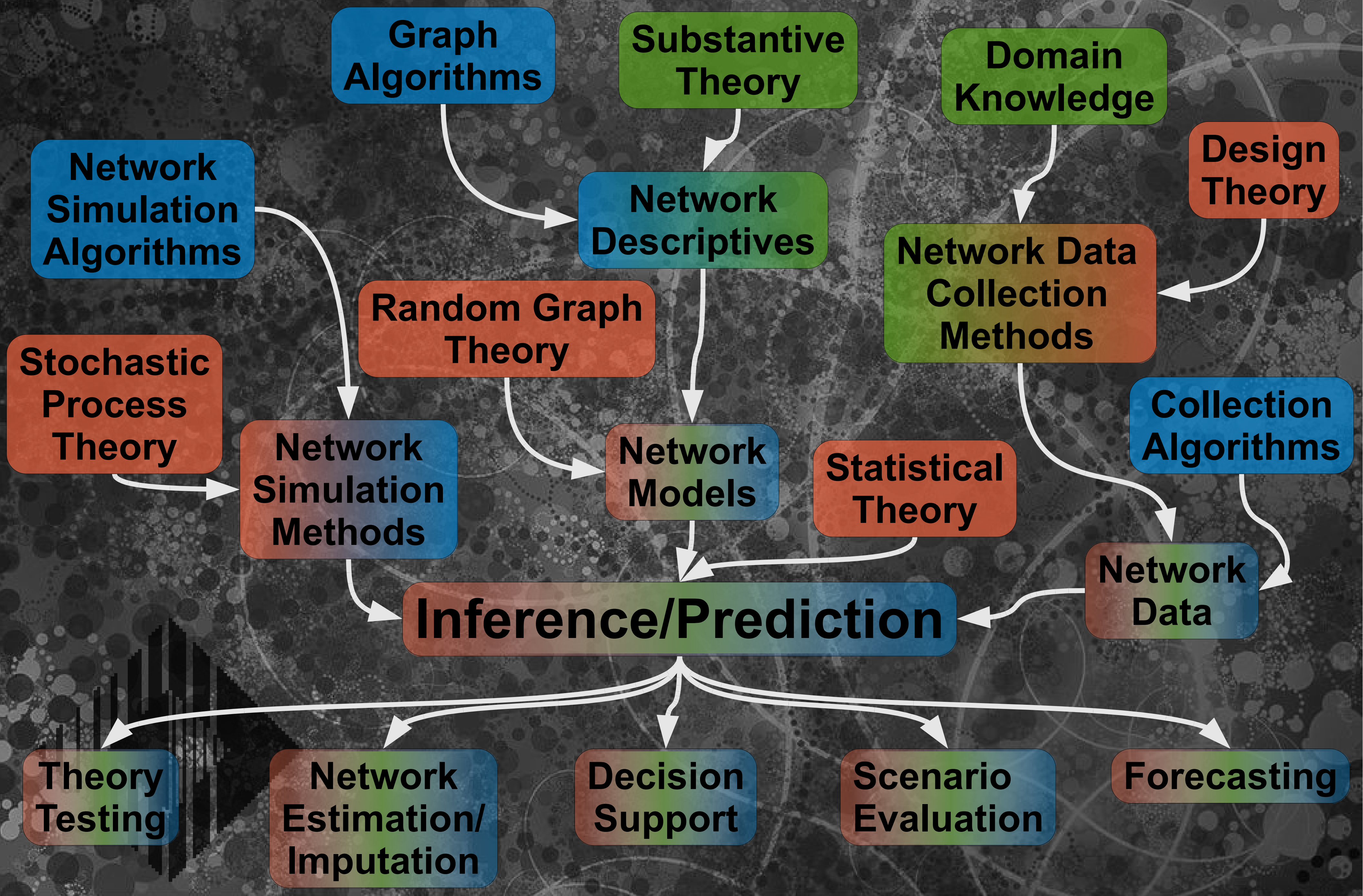
Mapping the Project Terrain



Mapping the Project Terrain



Mapping the Project Terrain

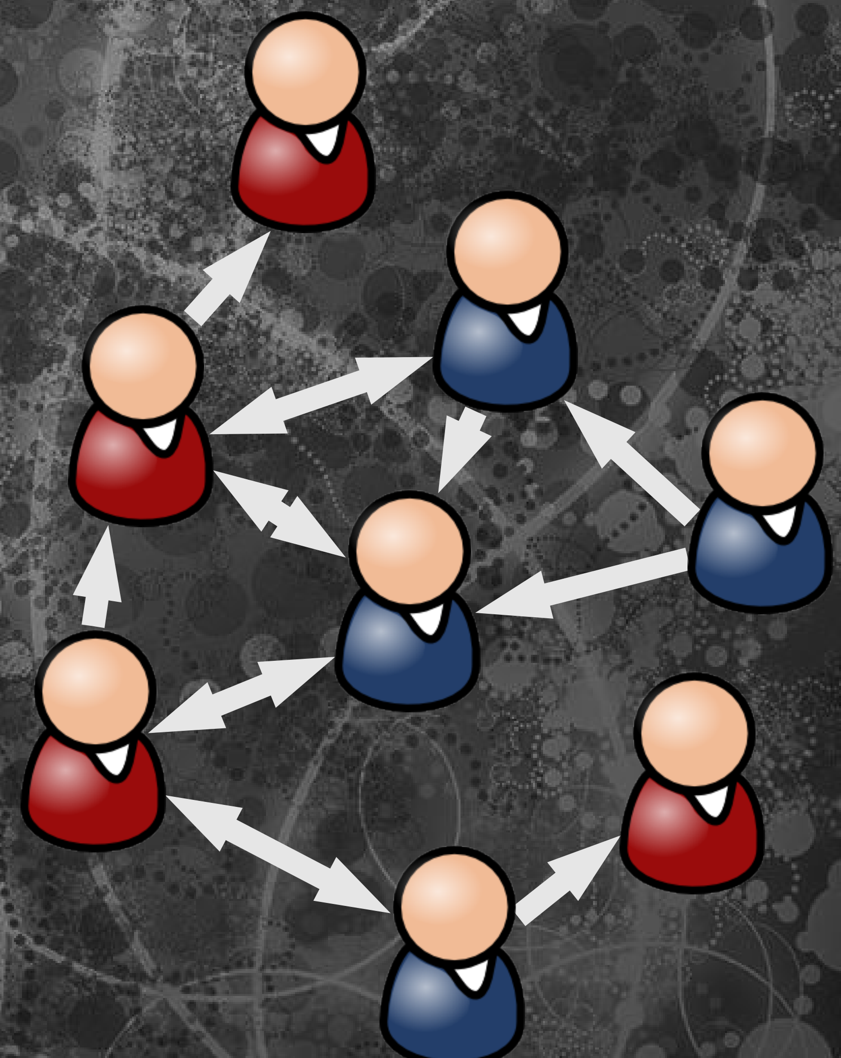


Why Statistical Models for Social (and Other) Networks?

- Social systems are complex
  - Many parts that affect each other
  - Substantial heterogeneity
- Many mechanisms involved
- We're not good at measuring them
  - Usually only see small chunks (and see above)
  - Error-prone observations
- Upshot: the network we see may result from many mechanisms, plus noise and unobserved factors
  - To draw conclusions about what is going on, must account for uncertainty
  - Predictions, conclusions should reflect this
  - Such goals require a statistical approach

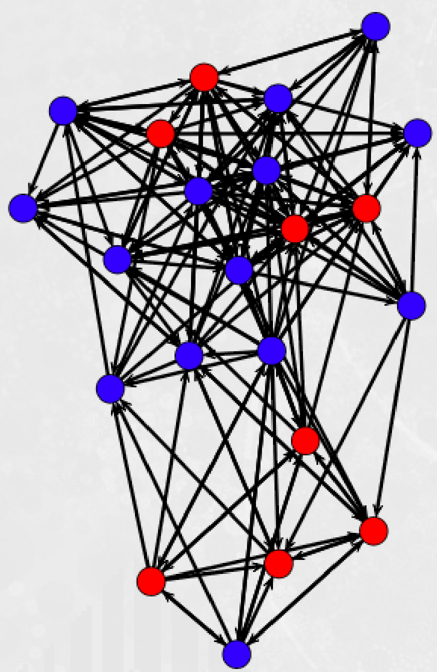
Motivating Example: The Reds and The Blues

- Consider a hypothetical community w/two groups - the "Reds" and the "Blues"
- Assume we are concerned with cooperation and trust in the community during a period of upheaval
- Our information is limited, but presume that we can observe networks of trust/friendship within representative subgroups....



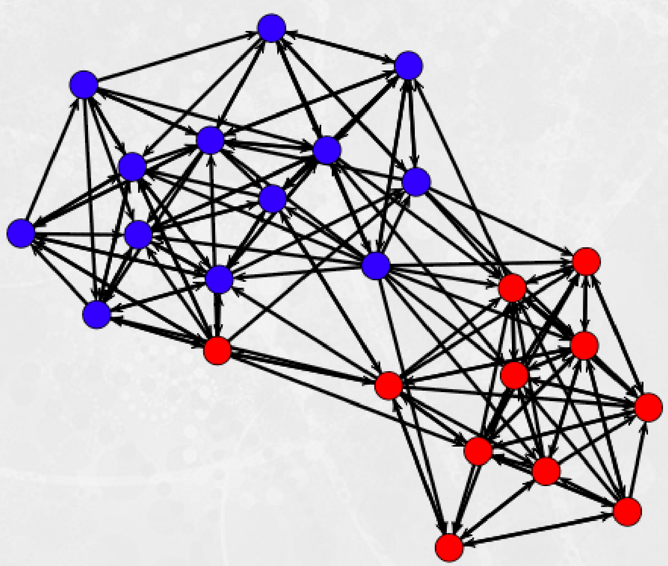
A Polarization Puzzle

Time 1



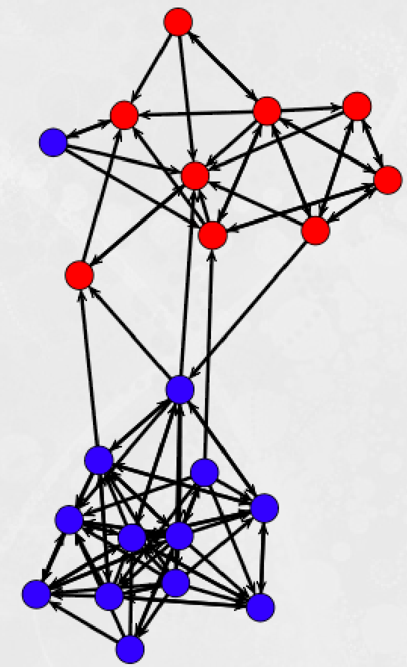
N=22

Time 2



N=24

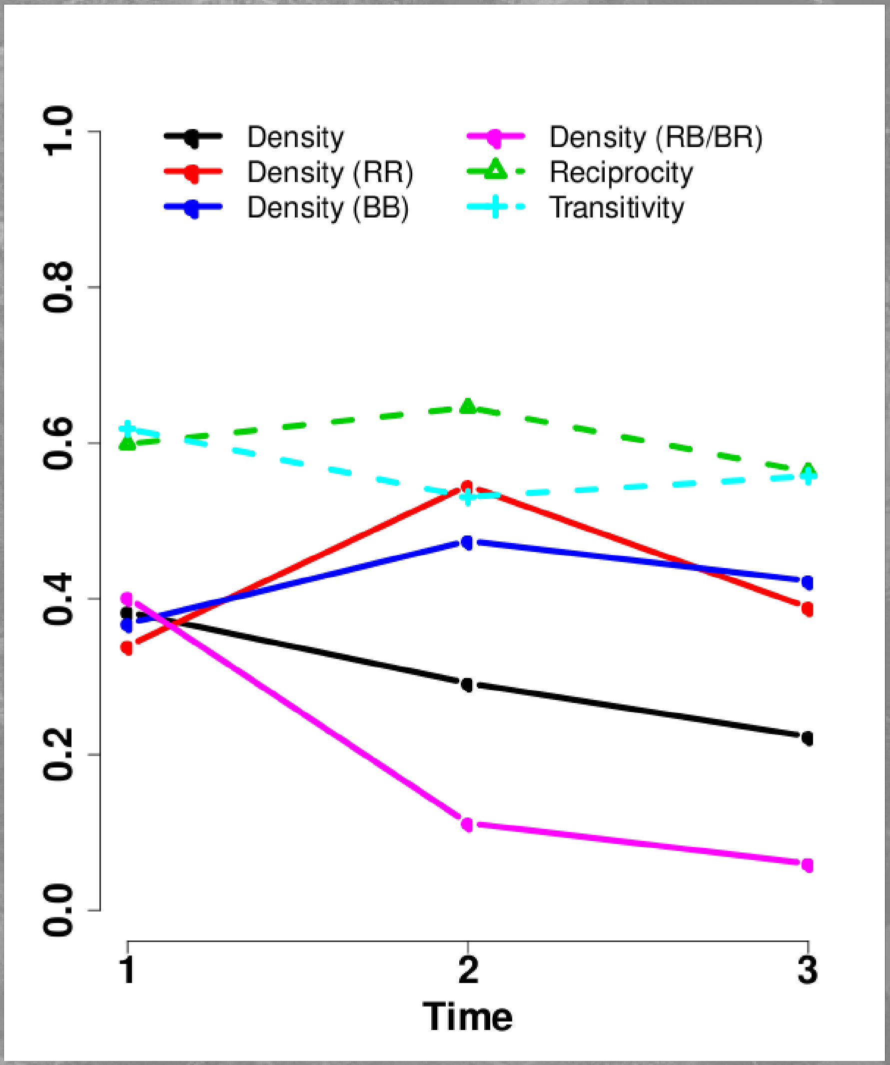
Time 3



N=22

First Step: Raw Descriptives

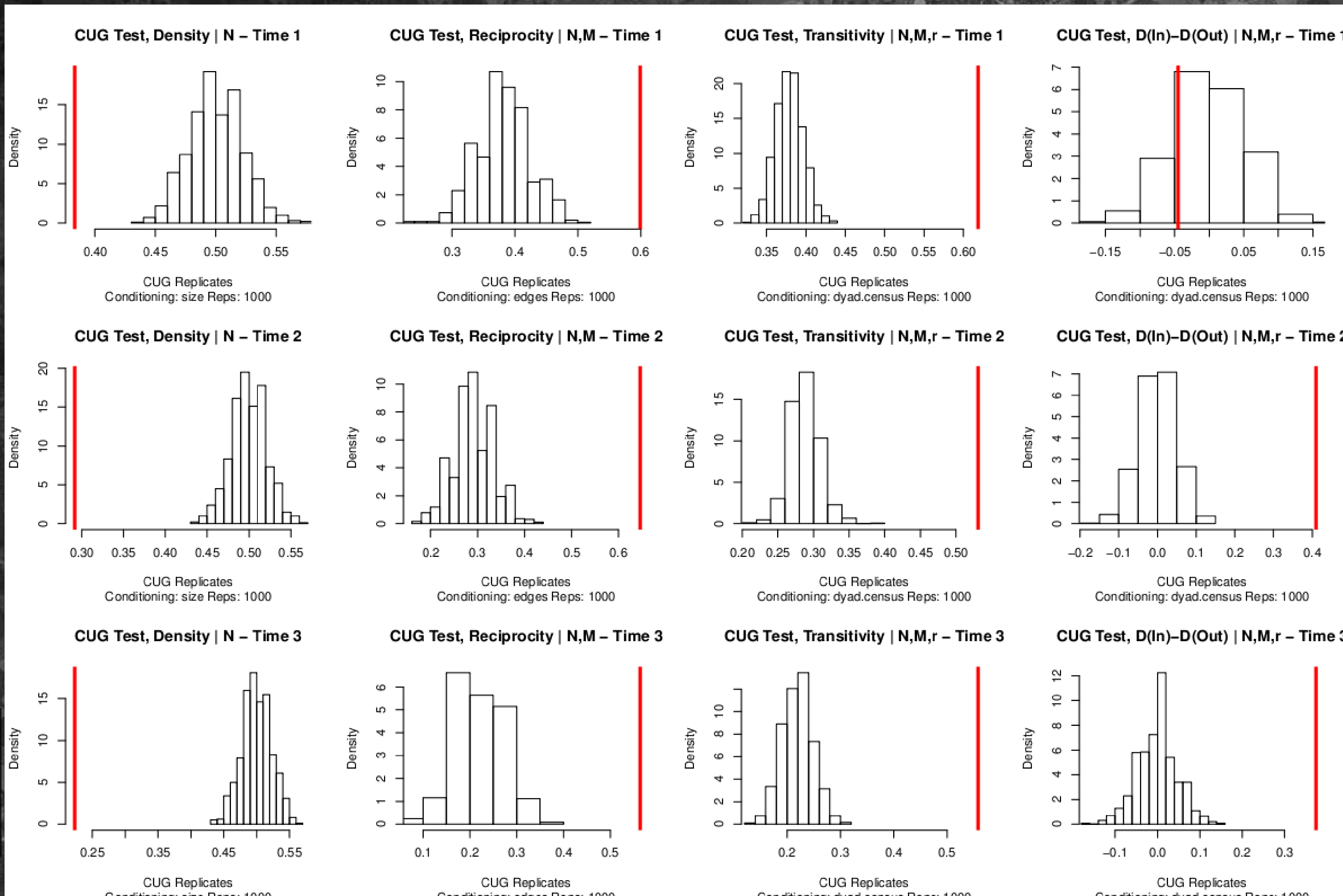
- Without a statistical approach, one is limited to description
- Here, some typical examples:
  - Density seems to fall slightly, although this masks an in/out-group difference
    - Red/Blue groups look similar
  - Moderately reciprocal, transitive networks, w/little change
- Gives a more precise accounting of events, but not very insightful
  - Are these changes even atypical of chance events?



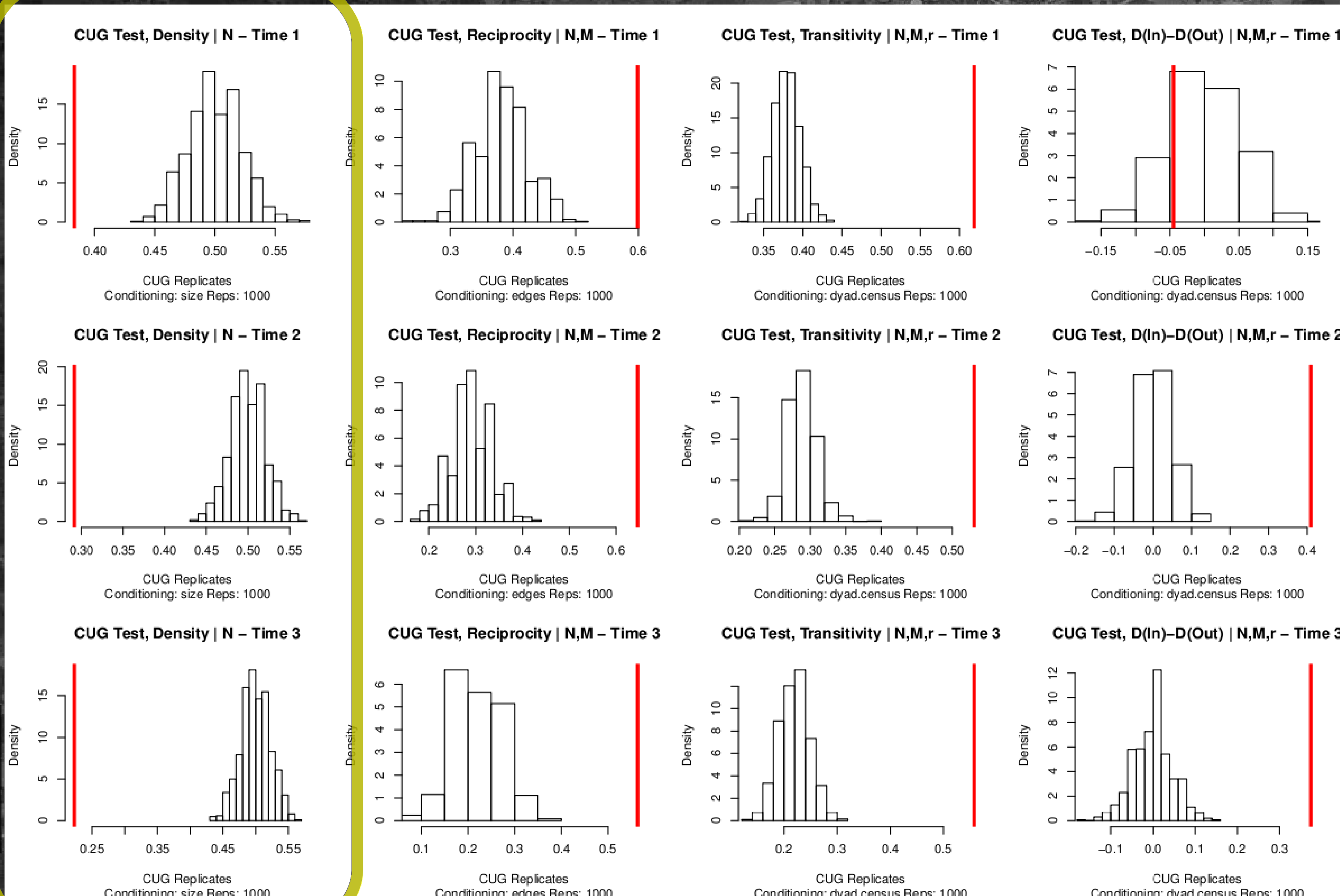
Next Step: Baseline Models

- Slight refinement: compare network properties to simple "baseline" models
  - E.g., uniform random graphs, conditional on a few properties
- Most elementary statistical approach
  - Assesses whether combinatorics + elementary properties are sufficient to account for observations
- Allows us to ask simple, marginal questions
  - Is density atypical of population of all graphs given N?
  - Is reciprocity atypical of graphs given N,M?
  - Are transitivity, difference in in-group/out-group densities atypical of graphs given N,M,r?
- Compare to classical null hypothesis testing

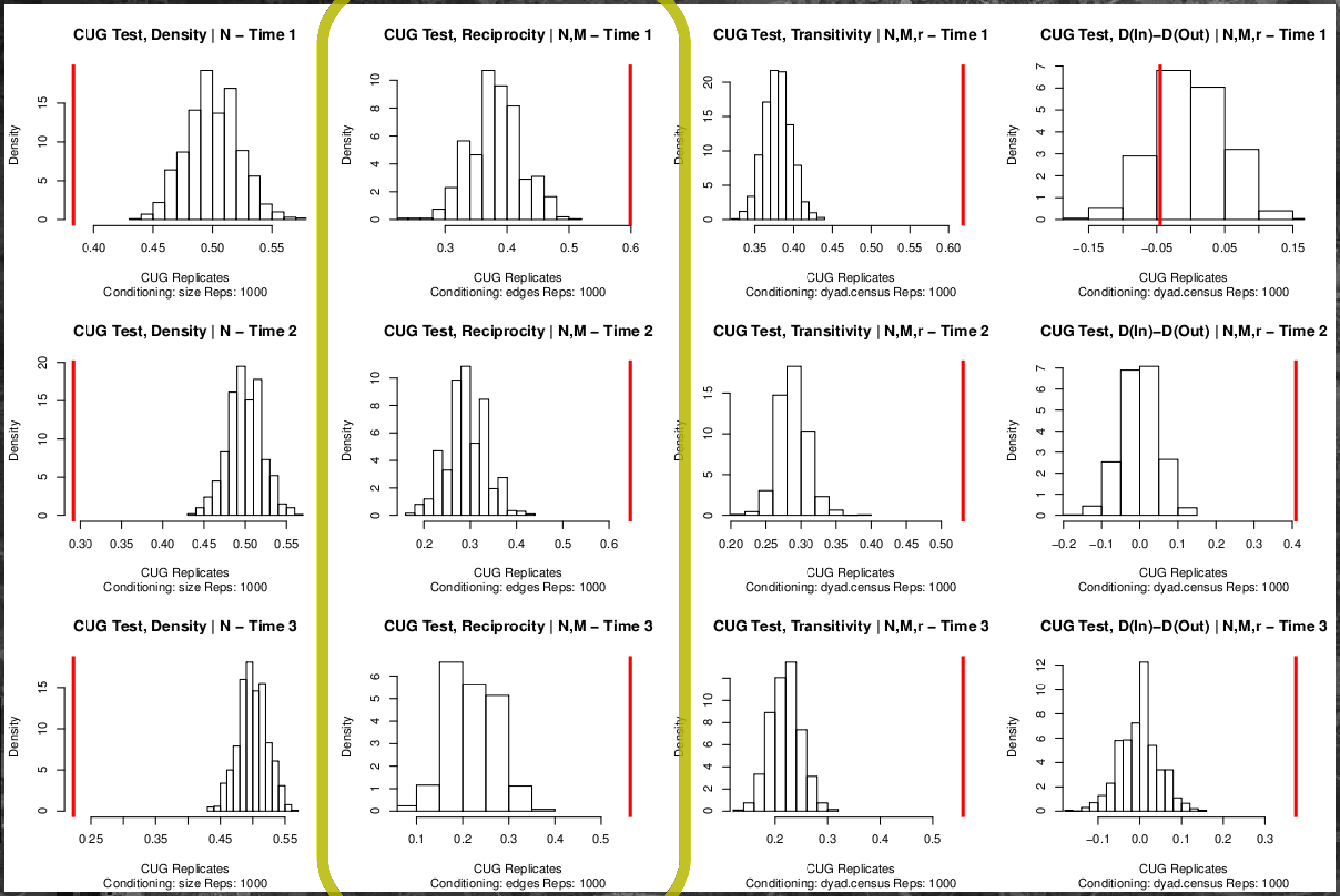
Baseline Comparisons



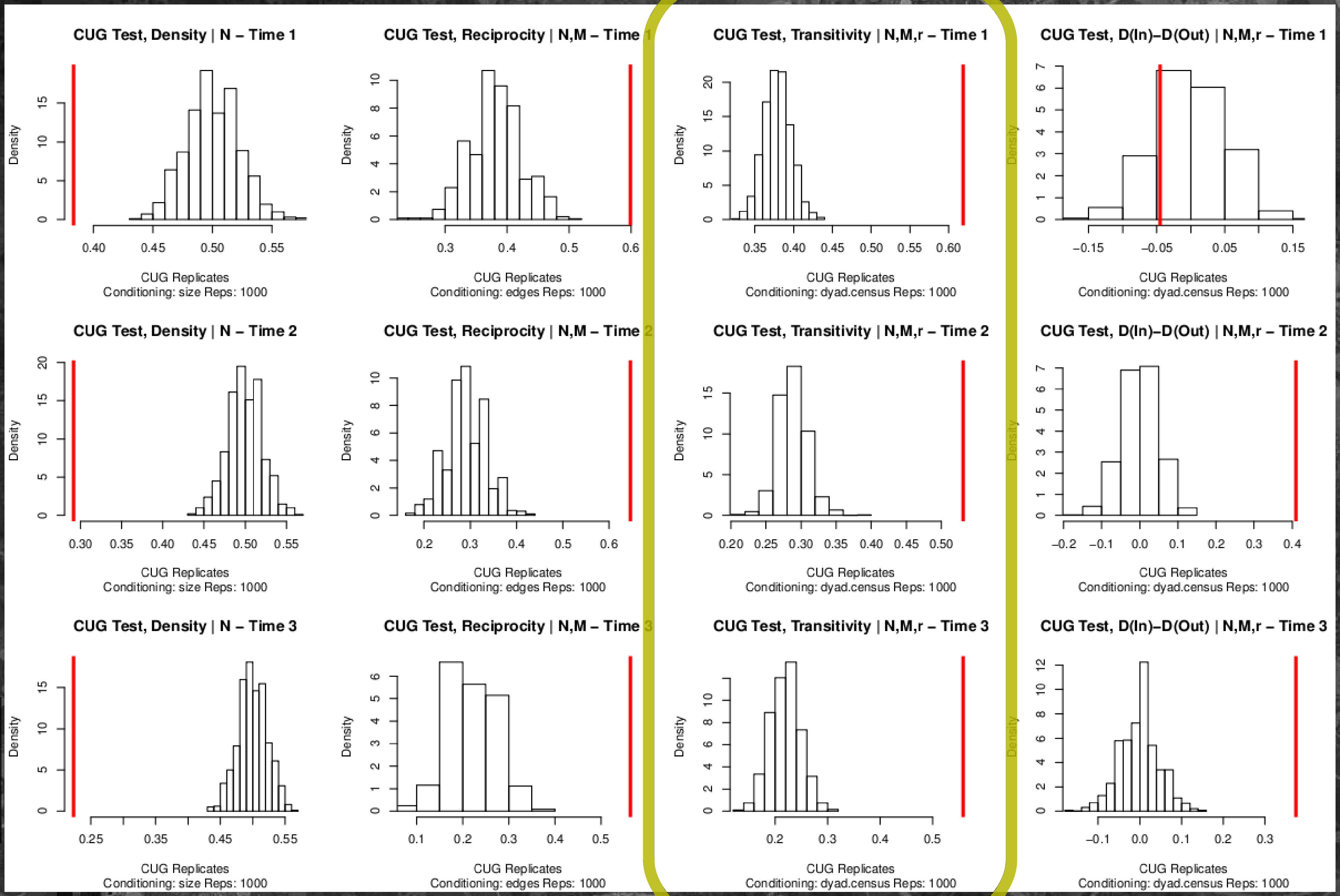
Baseline Comparisons



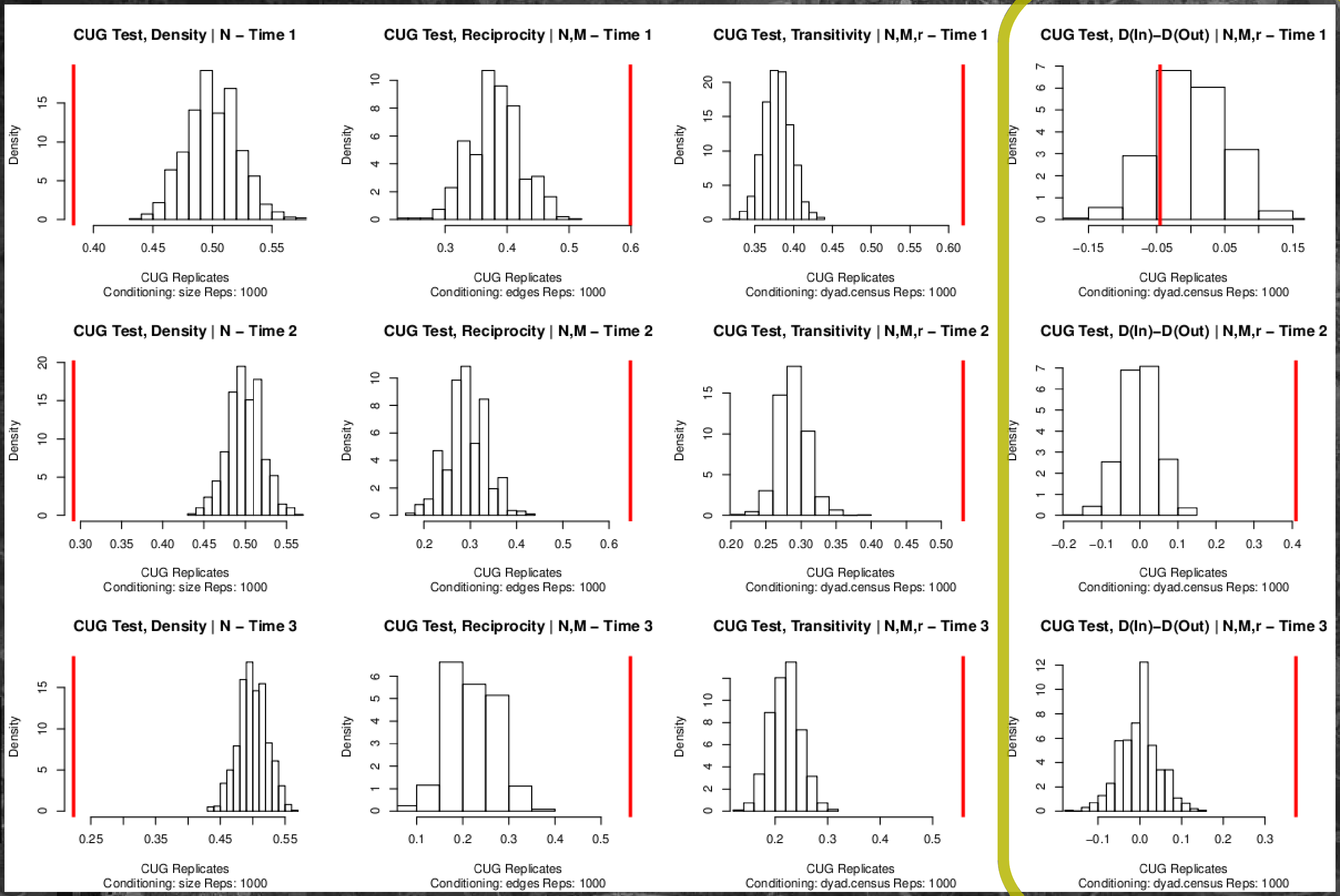
Baseline Comparisons



Baseline Comparisons



Baseline Comparisons

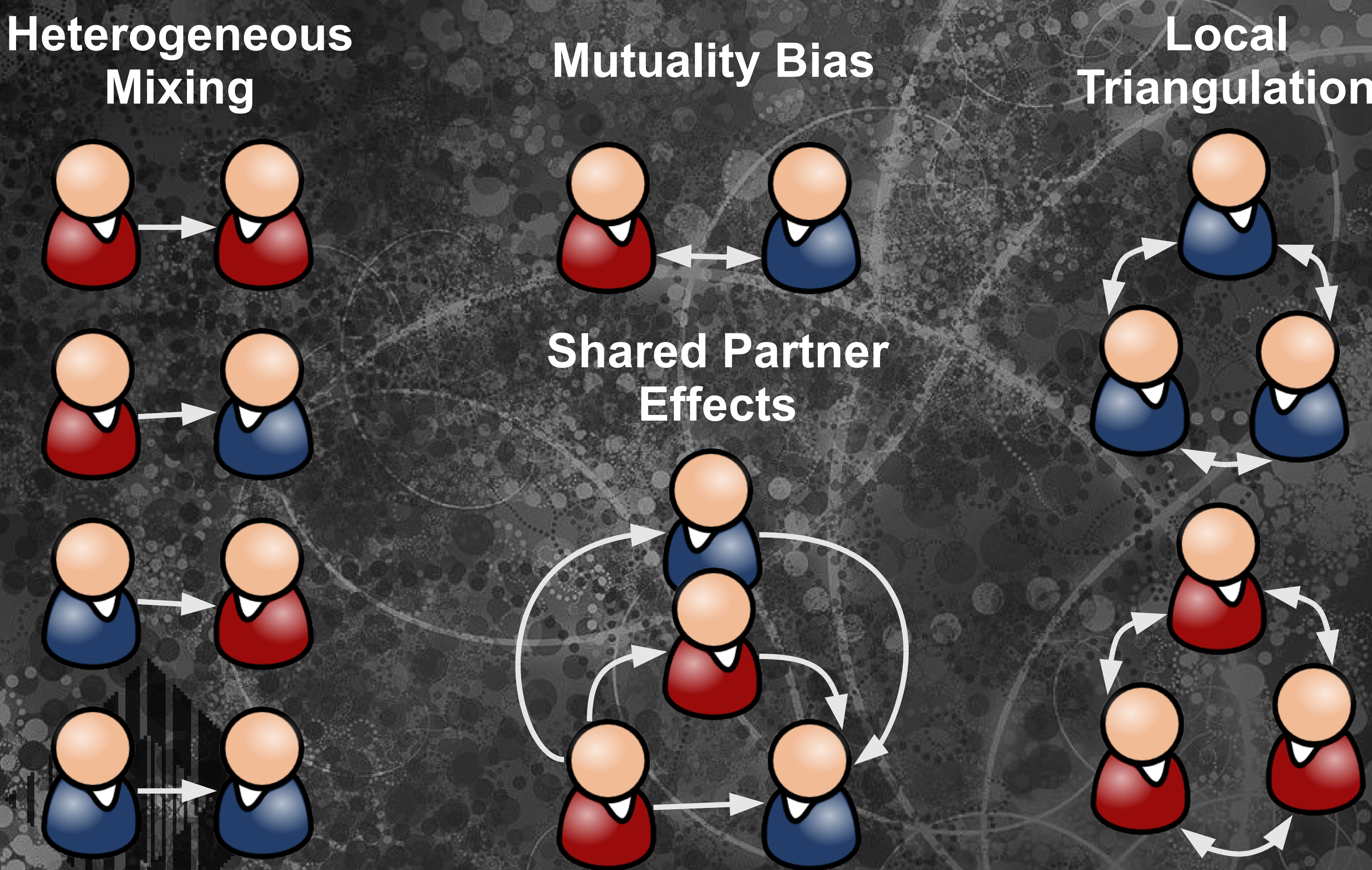


Beyond the Baselines

- Baseline models only take us so far
  - Few statistics lend themselves to conditioning
  - Difficult to look at multiple biases at once
  - Answers are qualitative in nature
  - Hard to account for sampling, error, etc.
  - Given "rejection" of the baseline, no clear path for further modeling

- Solution: parametric models
  - Identify candidate structural mechanisms
  - Parameterize using graph statistics
  - Fit models to data
    - Compare alternatives
    - Interpret parameter estimates
    - Assess adequacy
  - Can apply/extend for prediction, etc.

Sample Mechanisms



Evaluating Competing Explanations

Edges	Mixing	Mutuals	GWESP	LocalTri	AIC	Rank
1	0	0	0	0	1777.684	15
1	1	0	0	0	1565.073	14
1	0	1	0	0	1516.578	13
1	0	0	1	0	1227.656	2
1	0	0	0	1	1478.532	12
1	1	1	0	0	1428.158	11
1	1	0	1	0	1279.456	6
1	1	0	0	1	1416.441	10
1	0	1	1	0	1234.932	3
1	0	1	0	1	1348.794	9
1	0	0	1	1	1290.241	7
1	1	1	1	0	1216.762	1
1	1	1	0	1	1339.640	8
1	1	0	1	1	1238.285	5
1	0	1	1	1	1236.924	4

Evaluating Competing Explanations

Edges	Mixing	Mutuals	GWESP	LocalTri	AIC	Rank
1	0	0	0	0	1777.684	15
1	1	0	0	0	1565.073	14
1	0	1	0	0	1516.578	13
1	0	0	1	0	1227.656	2
1	0	0	0	1	1478.532	12
1	1	1	0	0	1428.158	11
1	1	0	1	0	1279.456	6
1	1	0	0	1	1416.441	10
1	0	1	1	0	1234.932	3
1	0	1	0	1	1348.794	9
1	0	0	1	1	1290.241	7
1	1	1	1	0	1216.762	1
1	1	1	0	1	1339.640	8
1	1	0	1	1	1238.285	5
1	0	1	1	1	1236.924	4

Evaluating Competing Explanations

Edges	Mixing	Mutuals	GWESP	LocalTri	AIC	Rank
1	0	0	0	0	1777.684	15
1	1	0	0	0	1565.073	14
1	0	1	0	0	1516.578	13
1	0	0	1	0	1227.656	2
1	0	0	0	1	1478.532	12
1	1	1	0	0	1428.158	11
1	1	0	1	0	1279.456	6
1	1	0	0	1	1416.441	10
1	0	1	1	0	1234.932	3
1	0	1	0	1	1348.794	9
1	0	0	1	1	1290.241	7
1	1	1	1	0	1216.762	1
1	1	1	0	1	1339.640	8
1	1	0	1	1	1238.285	5
1	0	1	1	1	1236.924	4

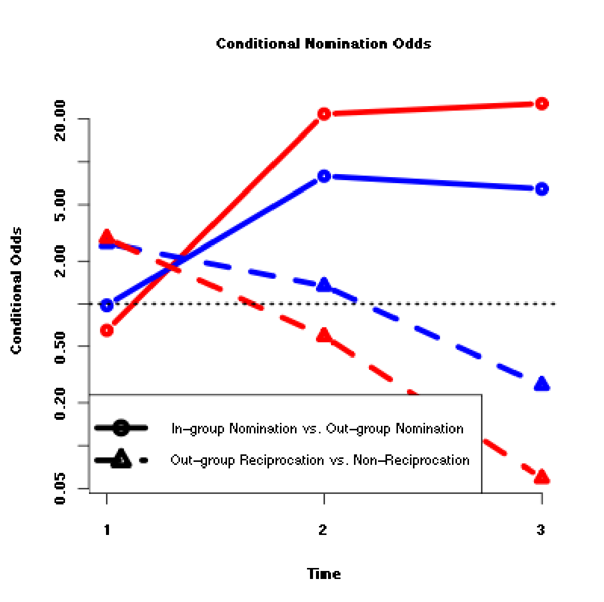
Evaluating Competing Explanations

Edges	Mixing	Mutuals	GWESP	LocalTri	AIC	Rank
1	0	0	0	0	1777.684	15
1	1	0	0	0	1565.073	14
1	0	1	0	0	1516.578	13
1	0	0	1	0	1227.656	2
1	0	0	0	1	1478.532	12
1	1	1	0	0	1428.158	11
1	1	0	1	0	1279.456	6
1	1	0	0	1	1416.441	10
1	0	1	1	0	1234.932	3
1	0	1	0	1	1348.794	9
1	0	0	1	1	1290.241	7
1	1	1	1	0	1216.762	1
1	1	1	0	1	1339.640	8
1	1	0	1	1	1238.285	5
1	0	1	1	1	1236.924	4

Interpreting the Mechanisms

	Time 1 MLE(SE)	Time 2 MLE(SE)	Time 3 MLE(SE)
Red→Red	-1.853 (0.291)	0.557 (0.226)	-1.069 (0.363)
Red→Blue	-1.421 (0.277)	-2.521 (0.428)	-4.317 (0.752)
Blue→Red	-1.501 (0.286)	-1.705 (0.354)	-2.809 (0.417)
Blue→Blue	-1.527 (0.198)	0.364 (0.226)	-0.948 (0.269)
Mutuals	2.484 (0.328)	1.992 (0.335)	1.489 (0.399)
GWESP	-0.030 (0.019)	-0.427 (0.031)	-0.018 (0.104)
GWESP(α)	1.218 (1.248)	0.744 (0.111)	0.598 (6.572)

- Sharp decline in out-group nomination propensity w/out systematic in-group shift
  - Conditional odds of in-group vs out-group nomination increase at time 2, stabilize
  - Effect somewhat stronger for Reds than Blues
- Decline in mutuality
  - Initially, both groups willing to conditionally reciprocate; by time 3, neither is!
- No clear trend in third-party effects
- Overall: out-group prefs, reciprocity key



And Beyond...

- Given an initial model family, there is much more one can do
  - Assess model adequacy versus target descriptives
  - Prediction (conditional, forecasting, scenario evaluation, etc.)
  - Extension/expansion given new data
- These are difficult or impossible using a purely descriptive framework (or even baseline models)

Looking Ahead

- Today's talks and posters will expand on these themes in various ways....
  - New methods for fitting network models
  - Algorithms to improve performance
  - New ways of parameterizing models
  - Applications to complex data sets
  - (and more!)
- Lots of work is in progress - don't hesitate to ask!