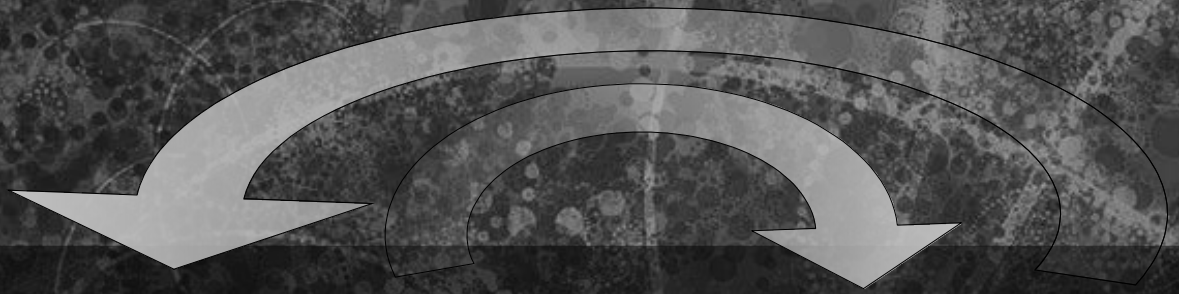
Advances in Scalable Modeling of Complex, Dynamic Networks

Carter T. Butts[12] and Zack Almquist[1]
[1]Department of Sociology
[2]Institute for Mathematical Behavioral Sciences
University of California, Irvine

Prepared for the November 12, 2010 UCI MURI AHM. This work was supported by DOD ONR award N00014-8-1-1015.

Advances in Scalable Modeling of Complex, Dynamic Networks



Zack Almquist[1] and Carter T. Butts[12]

[1]Department of Sociology

[2]Institute for Mathematical Behavioral Sciences

University of California, Irvine

Prepared for the November 12, 2010 UCI MURI AHM. This work was supported by DOD ONR award N00014-8-1-1015.

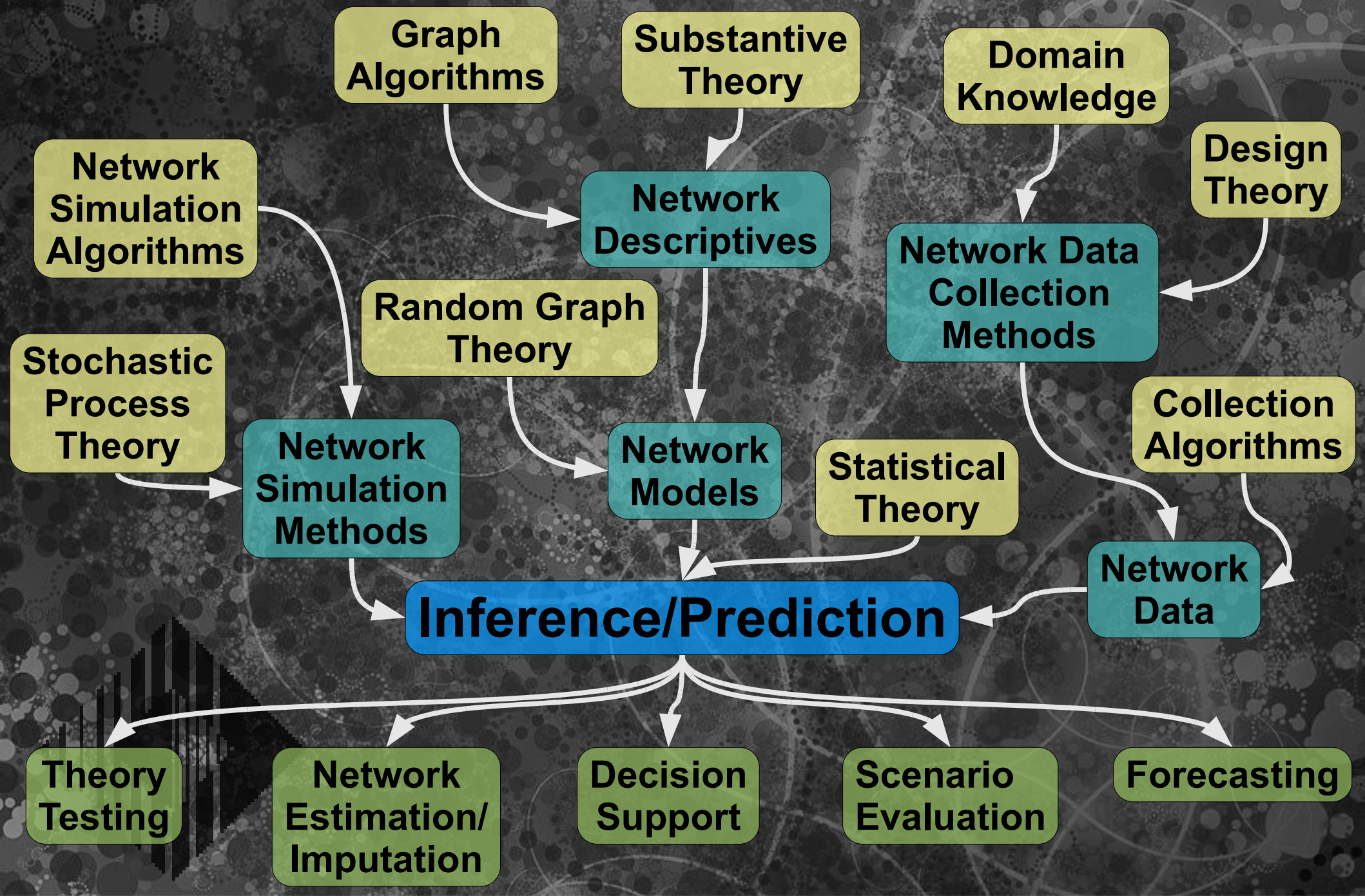
Advances in Scalable Modeling of Complex, Dynamic Networks

Zack Almquist[1] and Carter T. Butts[12]
[1]Department of Sociology
[2]Institute for Mathematical Behavioral Sciences
University of California, Irvine

(With input from the whole MURI team!)

Prepared for the November 12, 2010 UCI MURI AHM. This work was supported by DOD ONR award N00014-8-1-1015.

Dynamic Networks: Putting the Pieces Together



Modeling Dynamic Networks: Challenges and Advances

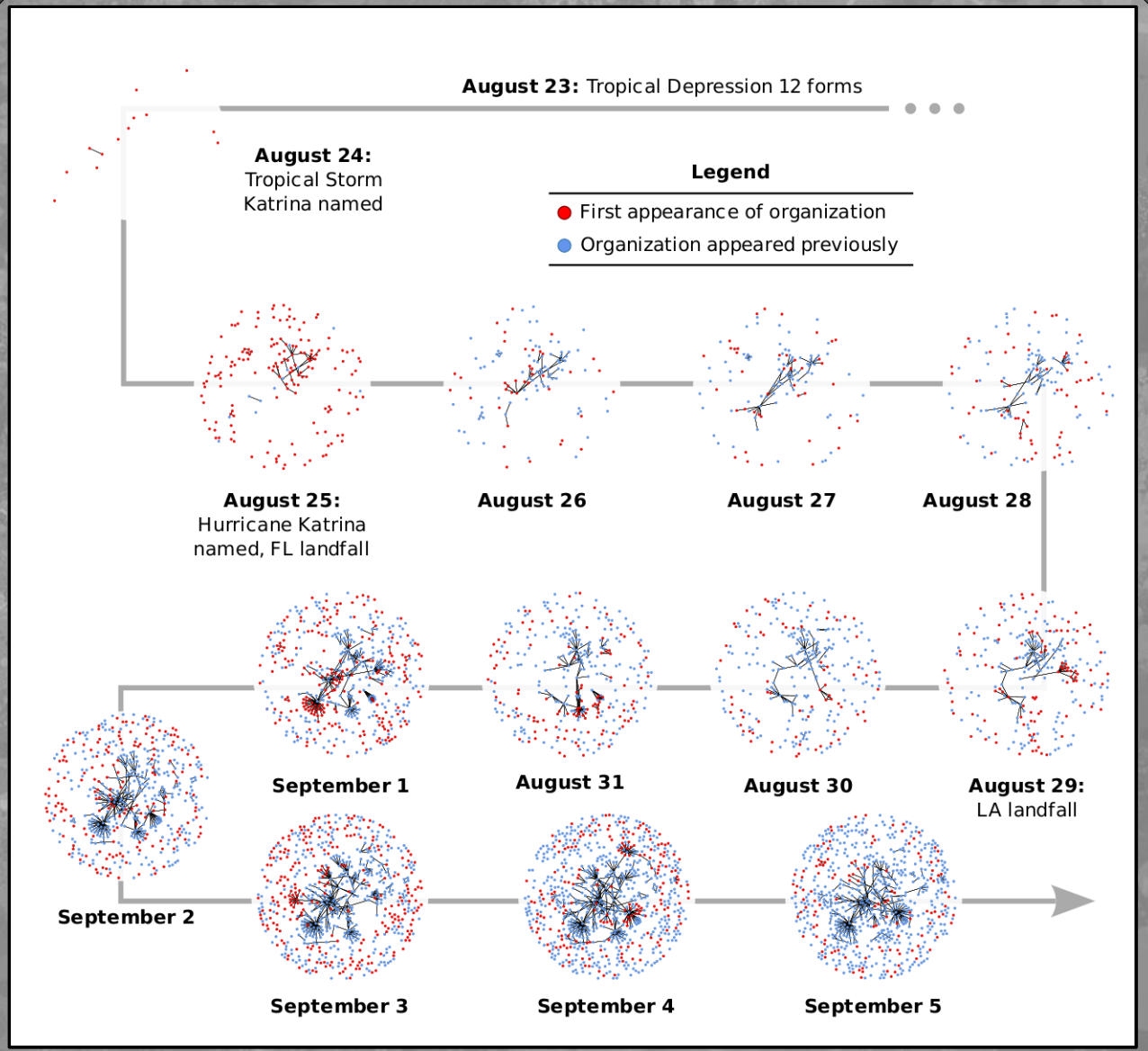
- Target area for our project, with many challenges:
  - Parameterizing models in a sensible and computable way
    - Models must reflect phenomenological understanding, but must also scale to real data
  - Making inference both principled *and* practical
    - Want accurate estimates, but can't wait forever for results
  - Dealing with rich, dynamic data
    - Real-world problems involve systems with complex covariates (e.g., geography, external events) that change over time
- Significant progress on several fronts
  - Panel data models, relational event models, latent structure models (see other talks, posters today!)
- This talk: some highlights and insights from one "thread" of this research

Scalability

- General problem: need to be able to model dynamic networks of reasonable size
  - Increasing number of data sets (including our own) with 100s/1000s of vertices, time points
  - Important for questions regarding large organizations, disasters, other complex settings
- At project start, primary approach (SIENA) limited to tens of vertices, fewer time points
- General temporal ERGM (TERGM) difficult, but there are workarounds (as we shall see)

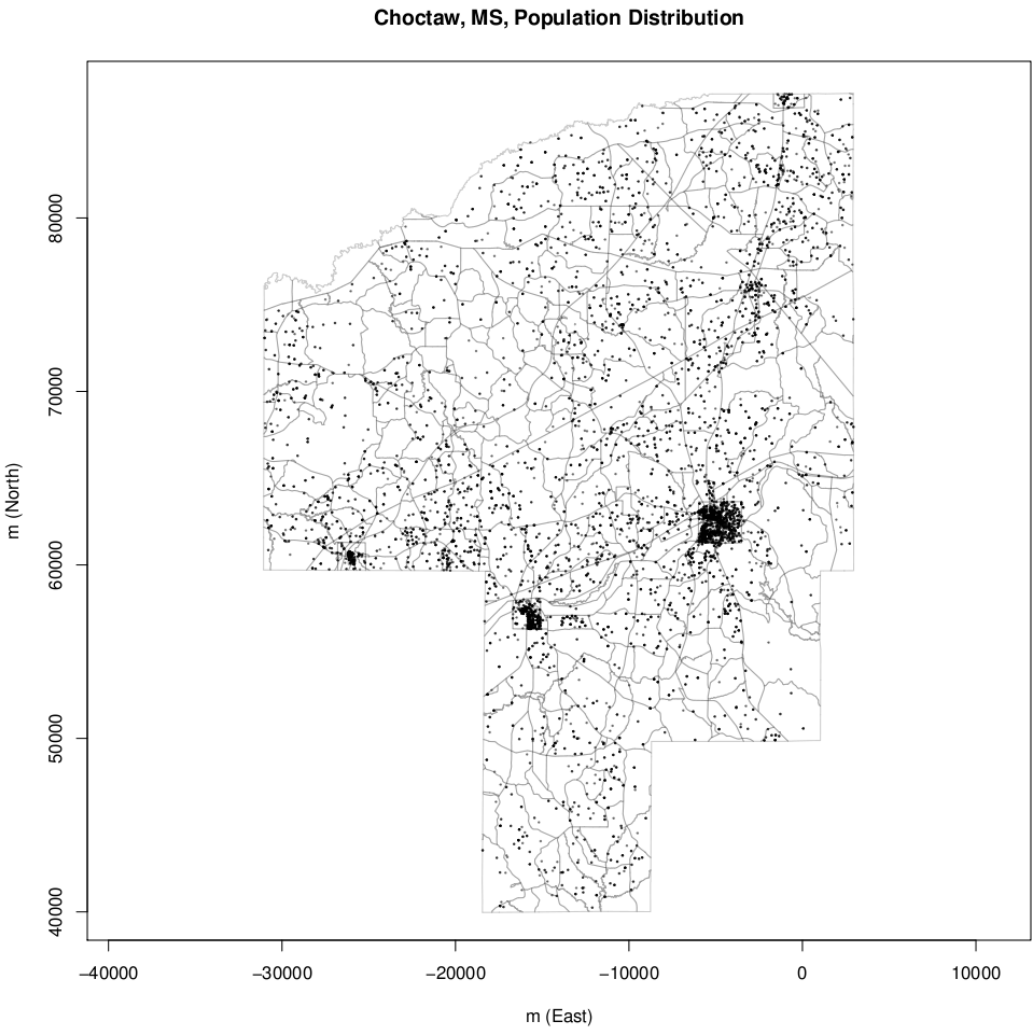
Vertex Set Dynamics

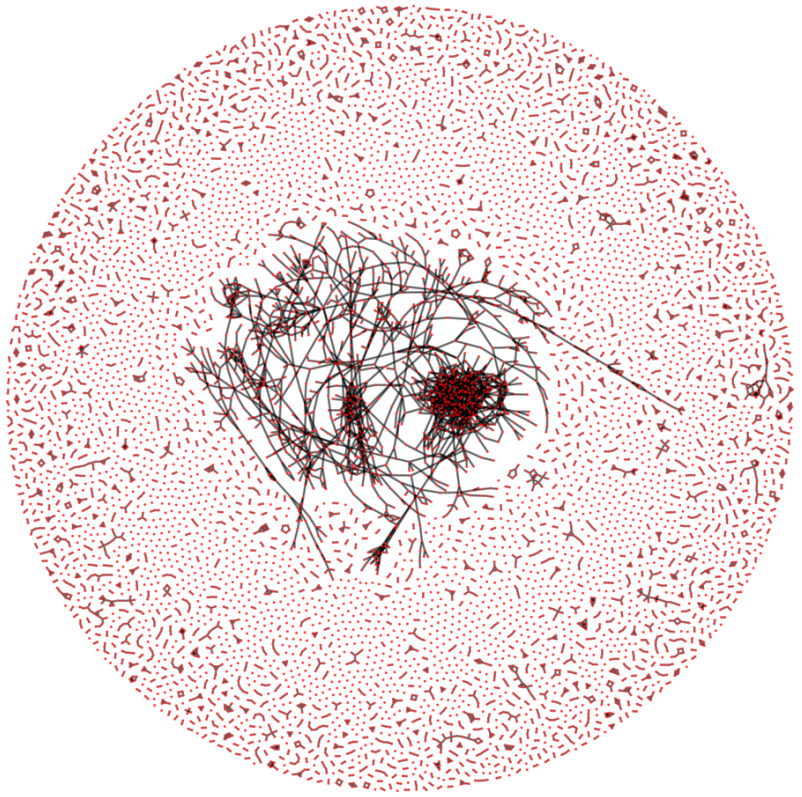
- Essentially overlooked problem in network literature: vertex set size, composition changes with time
- Extreme problem for phenomena like mass convergence in disasters, mortality, etc.
- More subtle issue when vertex covariates are present



Time-Evolution of the Hurricane Katrina EMON
(from Butts et al., 2010)

Context Effects: Spatial Context

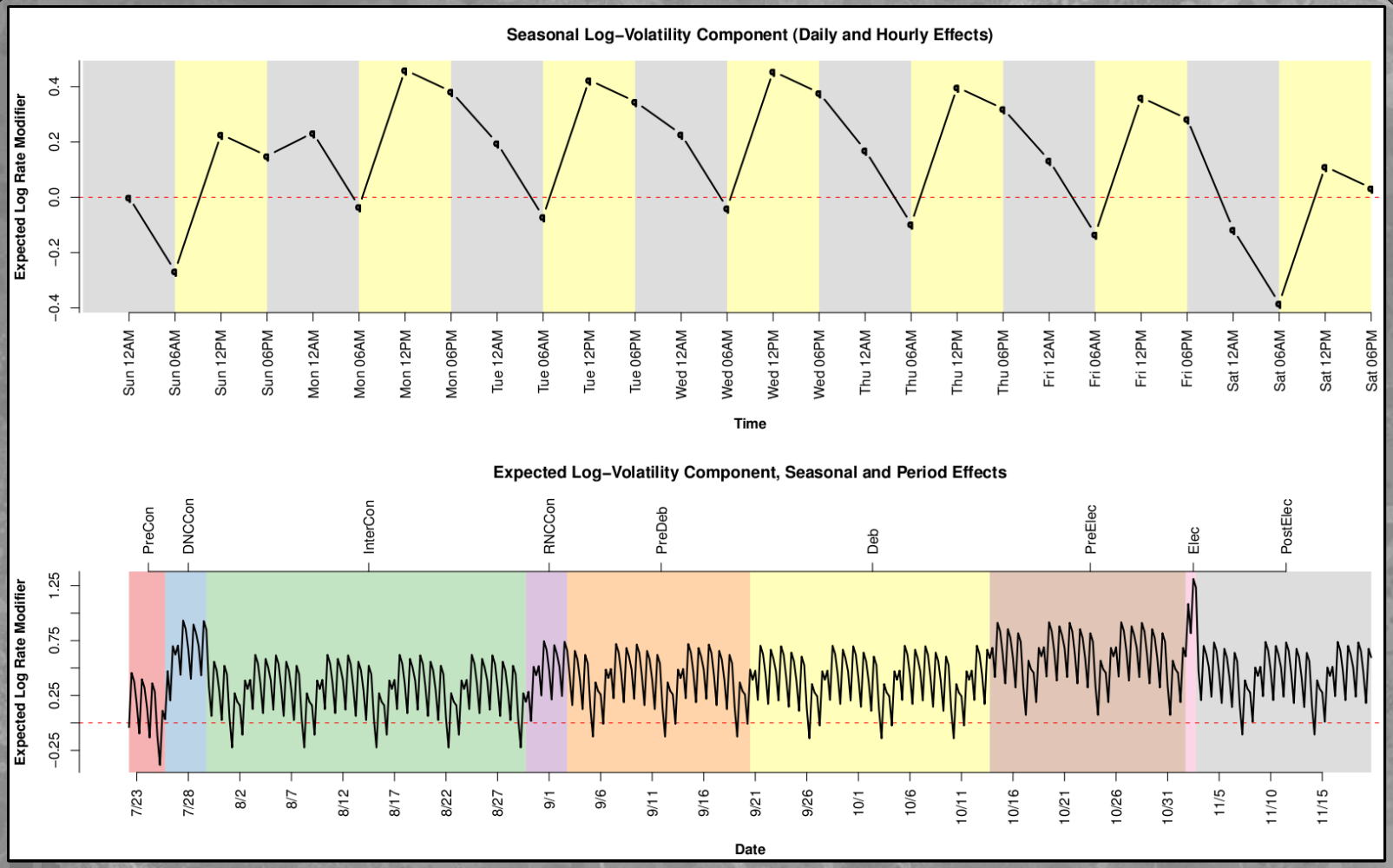




Choctaw, MS (N= 9,758) w/Spatial Bernoulli Realization (from Butts, 2010)

- Large-scale networks are usually geographically distributed
- Important to capture these effects (but computation a challenge)

Context Effects: Temporal Context



Seasonal and Period Effects in Volatility of English-Language Blog Networks, 2004 Electoral Cycle (from Butts and Cross, 2009)

- Human, organizational behavior shaped by seasonal mechanisms
- Need to capture seasonal, episodic influences on network evolution

Context Effects: External Forcing

- If all that weren't enough, networks experience external perturbations
- External forcing is itself dynamic; on large scales, geography can also matter
  - May not affect all vertices equally

Hurricane Katrina Storm Track
(from Butts et al., 2010)

Getting to Scalability: the Lagged Logistic Network Model

- Most serious obstacle to scalability: the ERG normalizing factor
- Approximate workaround: assume conditional dependence of edges, given the past
  - Introduced by Robins and Pattison (2001), extended by Hanneke and Xing (2007) and others
  - We extend existing practice by relaxing Markov assumption, adding vertex set dynamics
- Consequence: lagged logistic network model
  - Network at time t is modeled as logistic regression on prior states, along w/covariates
  - Allows us to leverage extensive machine learning literature on fast logistic regression for sparse matrices

Adding in Vertex Dynamics

- When the vertex set also evolves, we include it within the same framework
  - Support assumption: V_t is drawn from some maximal set V
  - Dependence assumptions: adjacency matrix Y_t depends on current vertex set, V_t; both $Z_t=(Y_t,V_t)$ can depend on past $Z_{t-k}=(Y_{t-k},V_{t-k})$ (generally up to some fixed $_k$)
  - Assuming that each $Y_{ij,t}$, $V_{i,t}$ are independent given the above leads to a joint logistic formulation:

$$\Pr(V_t|Z_{t-1},\ldots,Z_{t-k}) = \prod_{i=1}^{N} \text{logit}^{-1}\Big(w\big(I(v_i \in V_t), Z_{i-1},\ldots,Z_{i-k}\big)\Big)$$

$$\Pr(Y_t|V_t,Z_{t-1},\ldots,Z_{t-k}) = \prod_{(i,j)\in V_t\times V_t} \text{logit}^{-1}\Big(u\big(Y_{ij,t},V_t,Z_{i-1},\ldots,Z_{i-k}\big)\Big)$$

- Net result: flexible framework (TERGM subfamily) that readily scales to thousands of vertices/time points

Behavioral Interpretability: Stochastic Choice Dynamics

- Simple bridge from choice theory to network dynamics
- Core assumptions
  - Edge states unilaterally controlled (e.g., by sending actor)
  - Edge states may be reset at each time point
  - Edge state decisions made myopically, simultaneously, and in isolation; can depend upon past states or past/current environment, but not other decisions at same time point
  - Propensity to select a state (log odds) linear in expected utility difference to actor under conjectural variations
- Compare w/actor oriented framework of Snijders; we relax restriction on backward-looking behavior

Stochastic Choice Dynamics

- Above implies that $\Pr(Y_t = y_t) = \prod_{ij} \Pr(Y_{ij,t} = y_{ij,t})$
  - Y_t state of adjacency structure at time t
  - Probability can depend on any past state of Y, external covariates, etc.
- Behavioral rule implies logistic choice:

$$\begin{aligned}\operatorname{logit} \Pr(Y_{ij,t} = 1) &= \ln \frac{\Pr(Y_{ij,t} = 1)}{\Pr(Y_{ij,t} = 0)} \\ &= u_i\left(Y | Y_{ij,t} = 1\right) - u_i\left(Y | Y_{ij,t} = 0\right)\end{aligned}$$

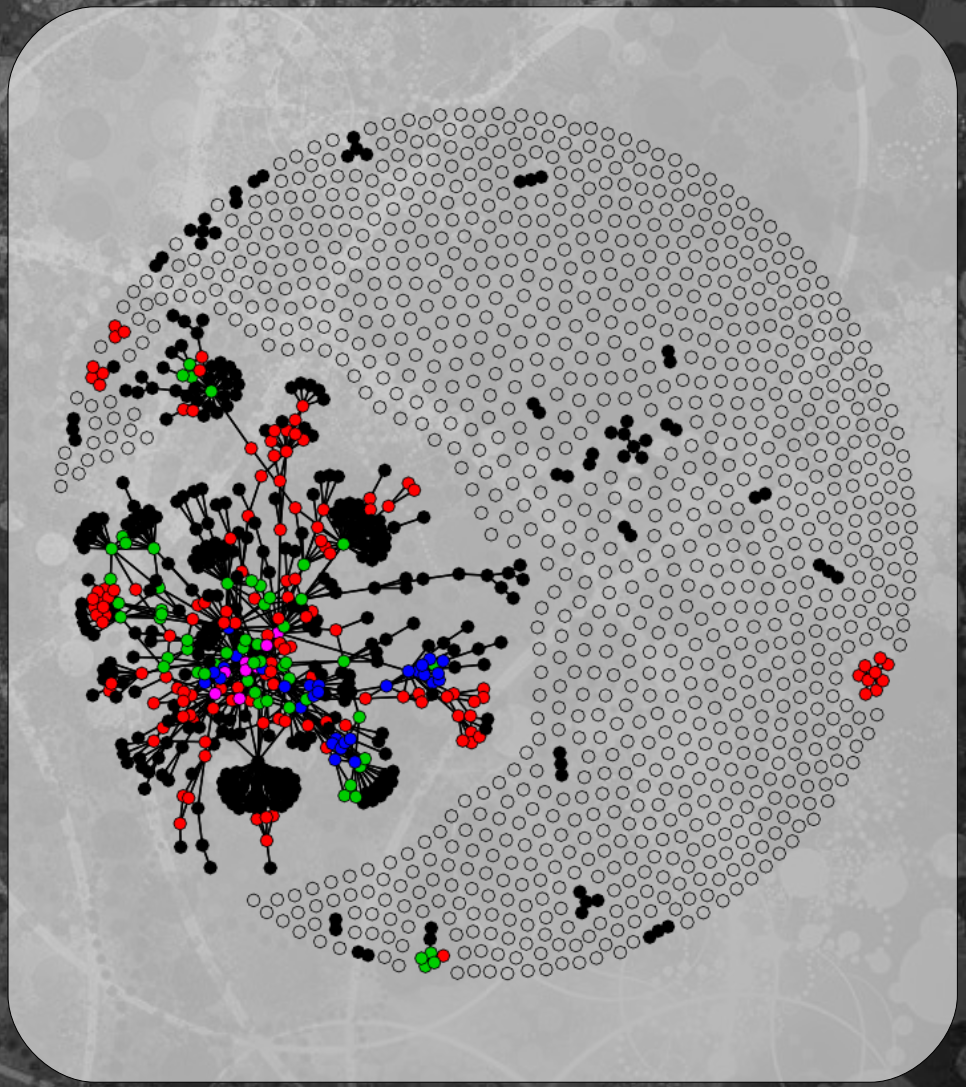
  - thus,

$$\Pr(Y_{ij,t} = 1) = \frac{\exp\left[u_i\left(Y | Y_{ij,t} = 1\right)\right]}{\exp\left[u_i\left(Y | Y_{ij,t} = 1\right)\right] + \exp\left[u_i\left(Y | Y_{ij,t} = 0\right)\right]}$$

  - which is the lagged logistic network model

Example Case: Hurricane Katrina

- Dynamic network of 1,577 organizations mobilized in the 13 days following storm formation
  - Daily snapshots; inclusion based on response activity, edges reflect collaboration on response tasks
- Network shows classic pattern of mass convergence
  - Grows from 13 to 775 organizations in a matter of days
- Geographical effects, extreme heterogeneity



Katrina Edge Model Highlights

Edge Parameter Estimates

	Model 1	Model 2	Model 3	Model 4	Model 5
BIC	45264.0197	32806.8854	32388.7429	31998.4955	31810.0173
Density	-8.735*	-8.6447*	-6.1638*	-5.7431*	-4.3685*
	(0.024)	(0.0288)	(0.0288)	(0.0292)	(0.0293)
Y_{t-1}		6.8045*	6.8713*	5.9812*	5.8815*
		(0.0601)	(0.0597)	(0.0627)	(0.0639)
$\log(n_{t-1})$			-0.4018*	-0.4998*	-0.5323*
			(0.0048)	(0.0049)	(0.0049)
Two-path			-0.009	-0.0859*	-0.1214*
			(0.0298)	(0.0292)	(0.0297)
Average Degree				0.1808*	0.1877*
				(0.0061)	(0.0061)
HQ State	2.4444*	1.8518*	1.7365*	1.6519*	1.2508*
	(0.0293)	(0.0392)	(0.0389)	(0.0396)	(0.04)
HQ City	0.8156*	0.6271*	0.6484*	0.6744*	-0.3382*
	(0.038)	(0.0544)	(0.0542)	(0.0549)	(0.055)
Fema Region	-0.3591*	-0.3191*	-0.3149*	-0.2196*	-0.3715*
	(0.028)	(0.0364)	(0.0362)	(0.0367)	(0.037)
Type	1.1564*	0.8226*	0.7919*	0.7923*	0.6179*
	(0.0263)	(0.0329)	(0.0328)	(0.0332)	(0.0335)
Scale	0.3641*	0.1904*	0.22*	0.1726*	0.0735
	(0.0324)	(0.0428)	(0.0426)	(0.0431)	(0.0436)
Lineage					1.9084*
					(0.1021)
Log Dist HQ city					-0.1539*
					(0.0056)

TABLE 1. Edge portion of Models 1 through 5 ranked by BIC score. Significance: '*' p-value < 0.05; z-test, under limiting assumptions, standard errors estimates based on resulting approximate EM algorithm of Gelman (2008) with a cauchy prior distribution centered at 0 with a scale parameter of 2.5.

Some evidence of preferential attachment, apparent coordinative brokerage (per Spiro)

Homophily by org type, but not by scale of operations; proximity in the "lineage" structure (Butts, 2009) important

Fairly strong propinquity effects from pre-disaster HQ location, despite emergent nature of the network

Katrina Vertex Model Highlights

Vertex Parameter Estimates

	Model 1	Model 2	Model 3	Model 4	Model 5
BIC	45264.0197	32806.8854	32388.7429	31998.4955	31810.0173
Intercept	-1.5749*	-2.3377*	-4.8207*	-4.7432*	-4.5078*
	(0.017)	(0.0202)	(0.0204)	(0.0204)	(0.0205)
Y_{t-1}		2.8074*	2.5377*	2.4185*	2.268*
		(0.035)	(0.0353)	(0.0356)	(0.0356)
$\log(n_{t-1})$			0.4547*	0.4525*	0.4273*
			(0.0035)	(0.0035)	(0.0035)
Degree				0.2025*	0.1989*
				(0.0309)	(0.0309)
HQ State	-0.1679*	-0.043*	-0.0724*	-0.0671*	-0.2274*
	(0.0171)	(0.0204)	(0.0205)	(0.0206)	(0.0206)
HQ City	0.403*	0.2786*	0.305*	0.2813*	0.3044*
	(0.0182)	(0.0217)	(0.0219)	(0.022)	(0.0221)
Fema Region	0.0558*	-0.0092	0.0174	0.0087	2.0954*
	(0.017)	(0.0203)	(0.0205)	(0.0205)	(0.0206)
Type	0.6832*	0.4285*	0.4687*	0.4724*	0.4519*
	(0.017)	(0.0203)	(0.0205)	(0.0206)	(0.0206)
Scale	-0.4681*	-0.2228*	-0.2668*	-0.3123*	-0.3264*
	(0.0171)	(0.0203)	(0.0205)	(0.0206)	(0.0206)
Sum of Lineage$_{t-1}$					-0.2943*
					(0.003)
Storm-track log Dist$_{t-1}$					0.0046*
					(0.001)

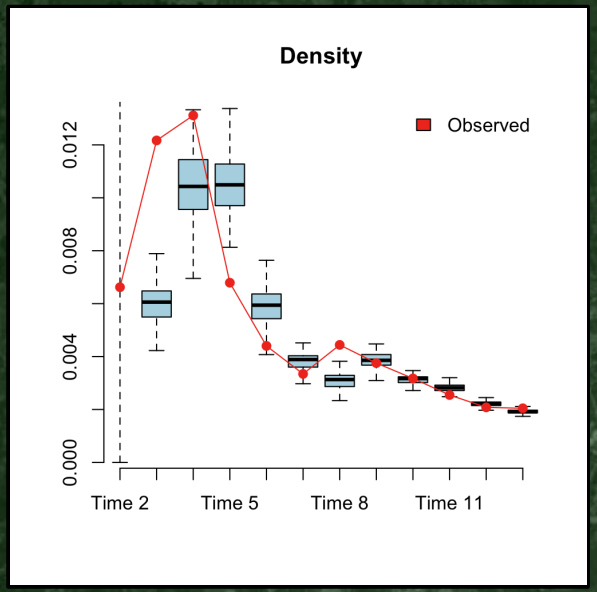
TABLE 1. Vertx portion of Models 1 through 5 ranked by BIC score. Significance: '*' p-value < 0.05; z-test, under limiting assumptions, standard errors estimates based on resulting approximate EM algorithm of Gelman (2008) with a cauchy prior distribution centered at 0 with a scale parameter of 2.5.

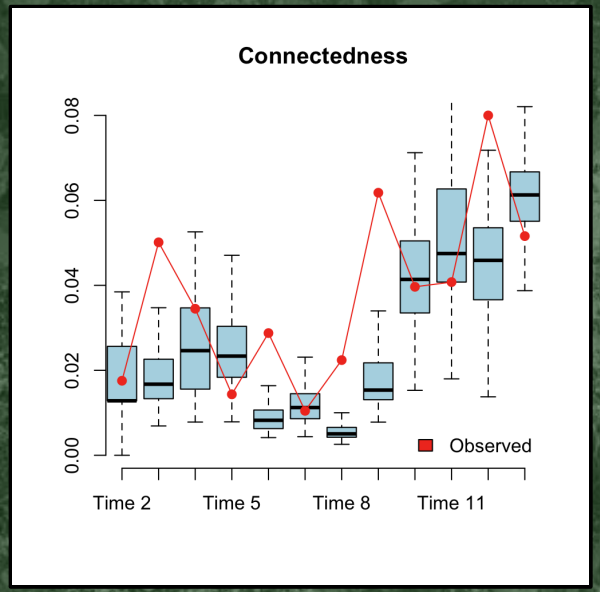
Overall positive feedback; orgs with more collaborations also more likely to remain mobilized

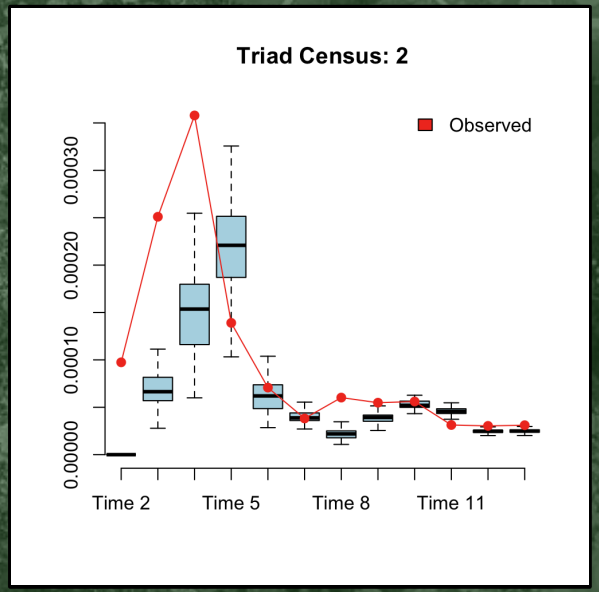
In general, the mobilization of other similar/ proximate orgs encourages mobilization (though scale/ state/lineage are heterophilous)

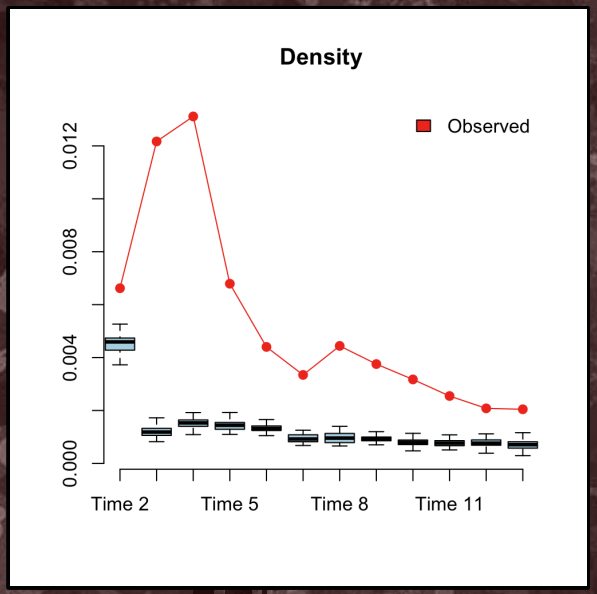
Mobilization slightly enhanced by being farther from storm track

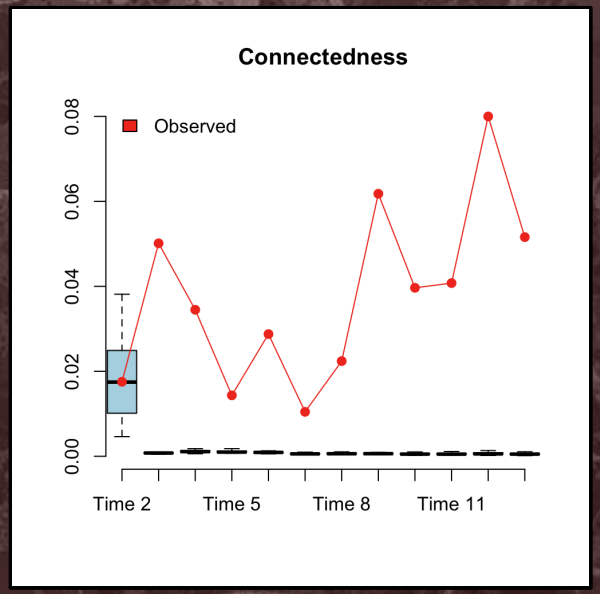
On Knowing Who Will Show Up

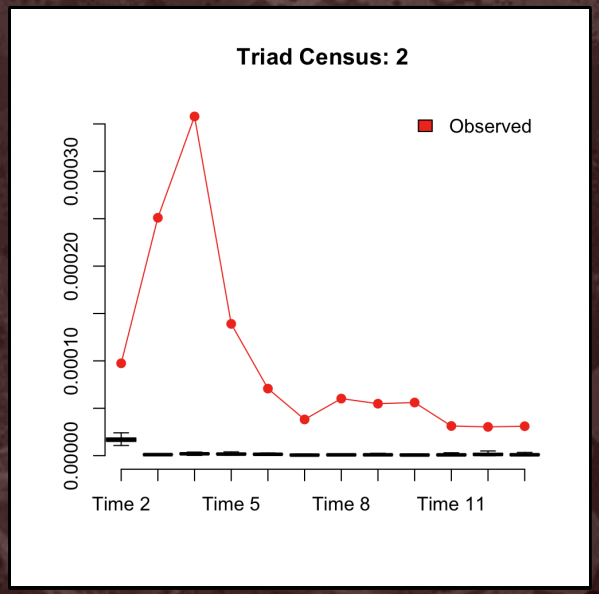




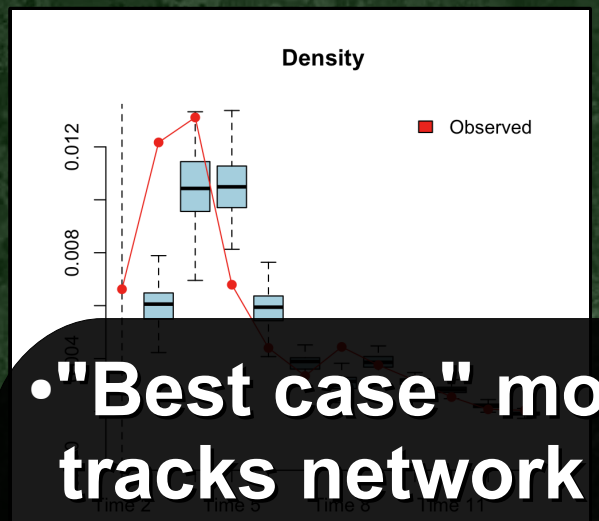


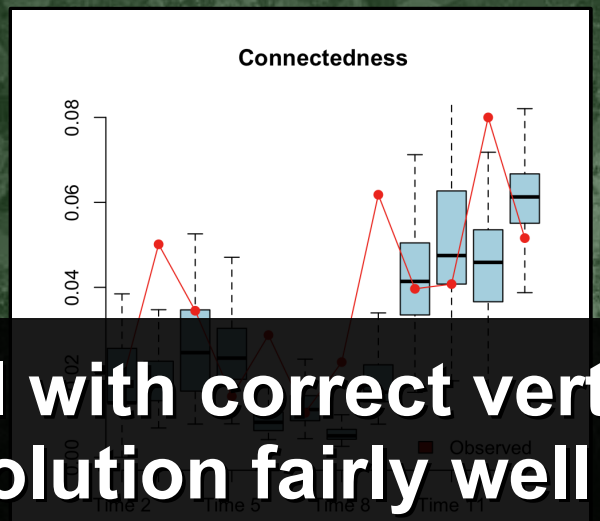


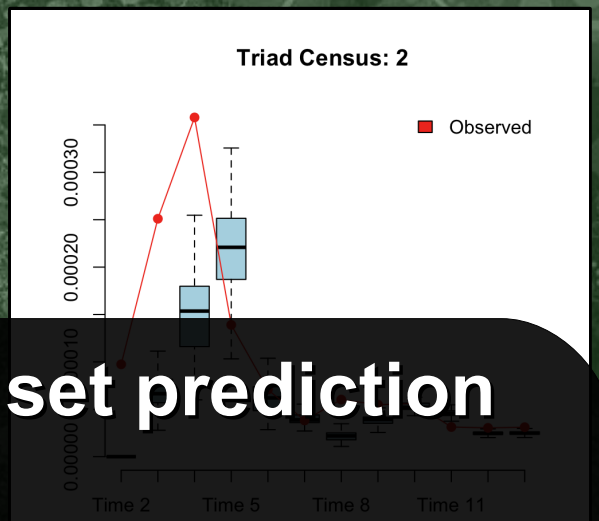




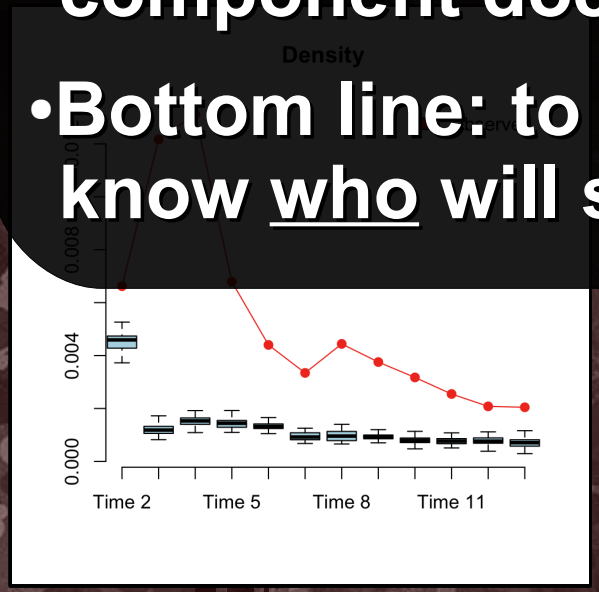
On Knowing Who Will Show Up

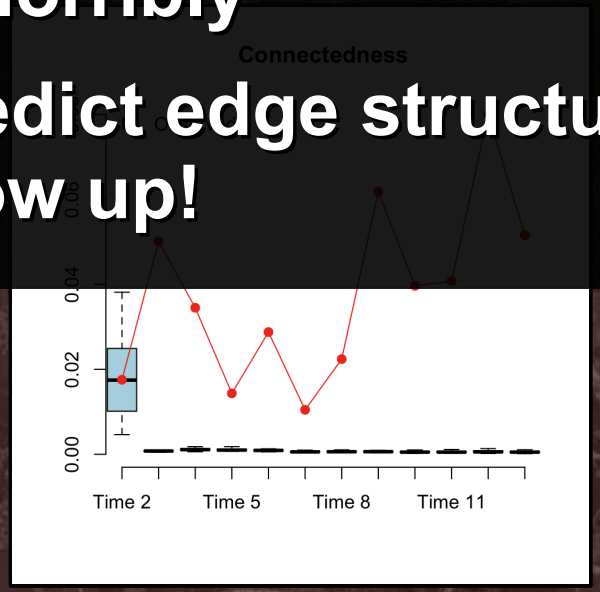


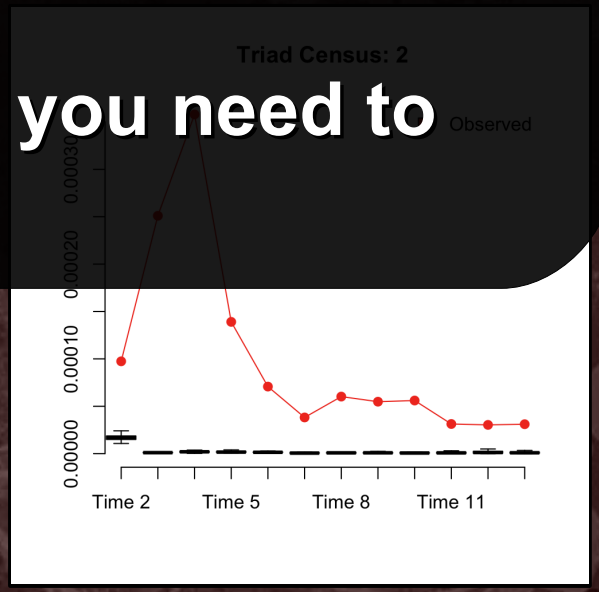




- "Best case" model with correct vertex set prediction tracks network evolution fairly well
- "Worst case" model with simplistic vertex set component does horribly
- Bottom line: to predict edge structure, you need to know <u>who</u> will show up!

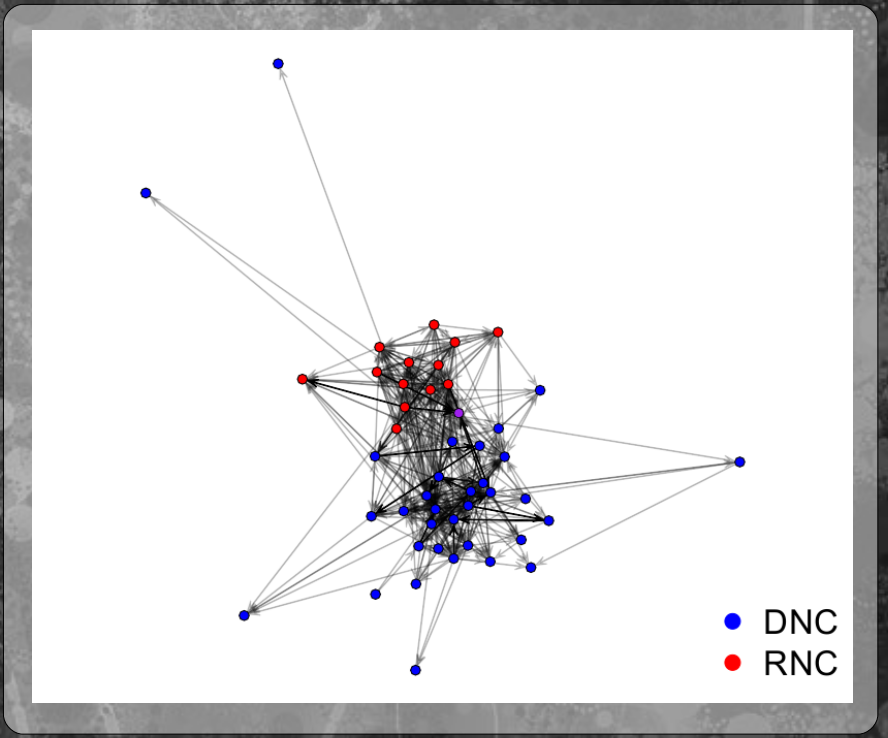






Example Case: DNC/RNC Blogs

- Interactions among partisan political blogs during the 2004 electoral cycle
- Dynamic network of 47 blogs over 121 day period
  - 34 credentialed for DNC, 14 for RNC (1 both)
  - Sampled 4 times daily, for 484 time points
- Example of interaction among contending parties
- Opportunity to study seasonal, period effects on behavior



Blog Network Edge Model: Highlights

DNC–RNC Blog Network

	Estimate	St. Error	z value	P(>|z|)
DNC	-4.52	0.02	-191.79	0.00
RNC	-3.21	0.03	-92.63	0.00
DNC→RNC	-5.41	0.06	-93.09	0.00
RNC→DNC	-4.50	0.04	-119.28	0.00
Y_{t-1}	10.32	0.03	396.08	0.00
Clique	0.34	0.03	11.08	0.00
Reciever	0.05	0.00	19.15	0.00
Sender	-0.02	0.00	-9.27	0.00
Group-2-path	0.21	0.01	22.45	0.00
Coss-Group-2-path	0.33	0.03	10.70	0.00
Group-Reciprocity	-0.51	0.05	-10.61	0.00
Between-Group-Reciprocity	0.63	0.22	2.91	0.00
θ_{D1}	0.10	0.02	4.91	0.00
θ_{D2}	-0.16	0.03	-5.43	0.00
H_2	-0.12	0.03	-3.72	0.00
H_3	-0.35	0.04	-9.71	0.00
H_4	-0.47	0.04	-13.01	0.00
Indegree×H_2	-0.12	0.05	-2.28	0.02
Indegree×H_3	-0.18	0.06	-3.28	0.00
Indegree×H_4	-0.11	0.06	-1.80	0.07
$Y_{t-1} \times H_2$	0.20	0.04	4.81	0.00
$Y_{t-1} \times H_3$	-0.30	0.05	-6.18	0.00
$Y_{t-1} \times H_4$	0.07	0.05	1.35	0.18
DNCCon	-1.92	0.10	-19.94	0.00
InterCon	-1.95	0.04	-55.24	0.00
RNCCon	-1.95	0.10	-20.34	0.00
PreDeb	-1.82	0.05	-40.28	0.00
Deb	-1.89	0.04	-47.54	0.00
PreElec	-1.93	0.05	-42.07	0.00
Elec	-2.42	0.19	-12.97	0.00
PostElec	-1.95	0.05	-40.26	0.00

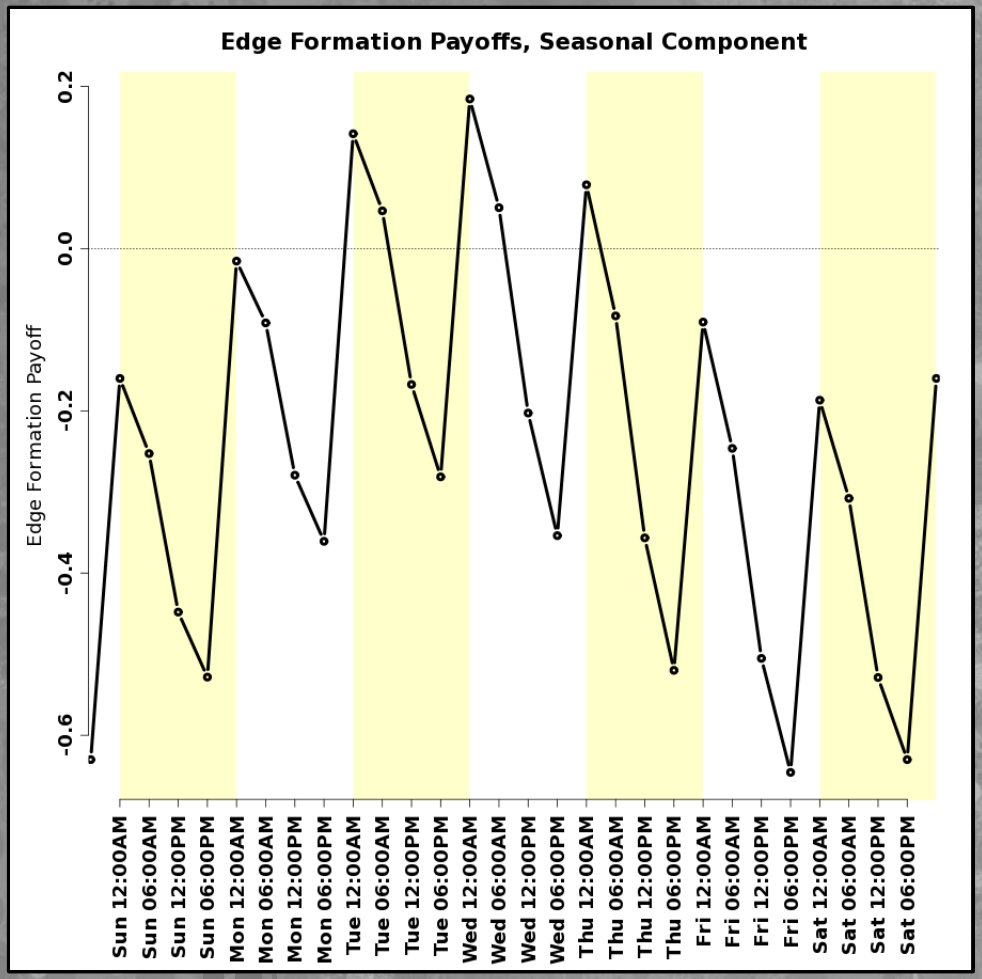
Payoff to ingroup citations higher than cross-group citations (but less true for RNC on a per-tie basis)

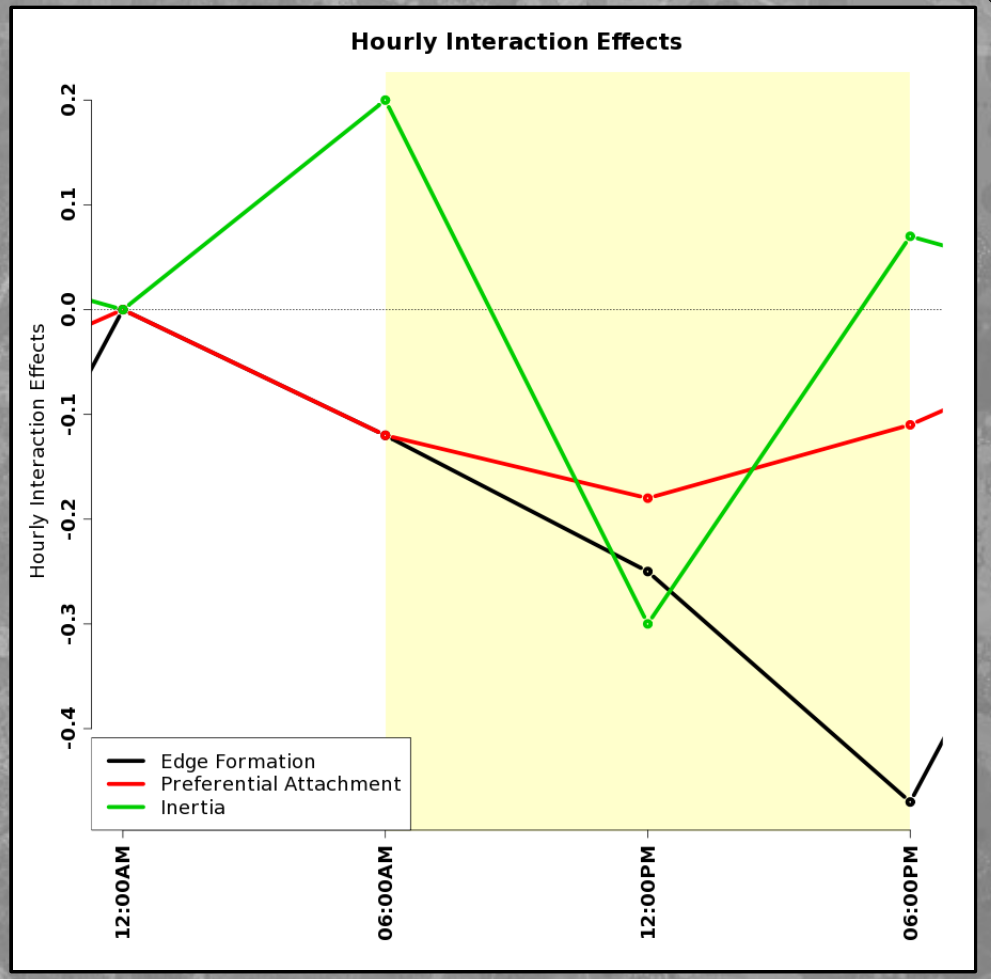
Clique and two-path embeddedness act as citation incentives; tend to cite others who are widely cited, but not to cite those with many out-citations

Reciprocation pays off, but not in one's own group! (Deference vs conflict)

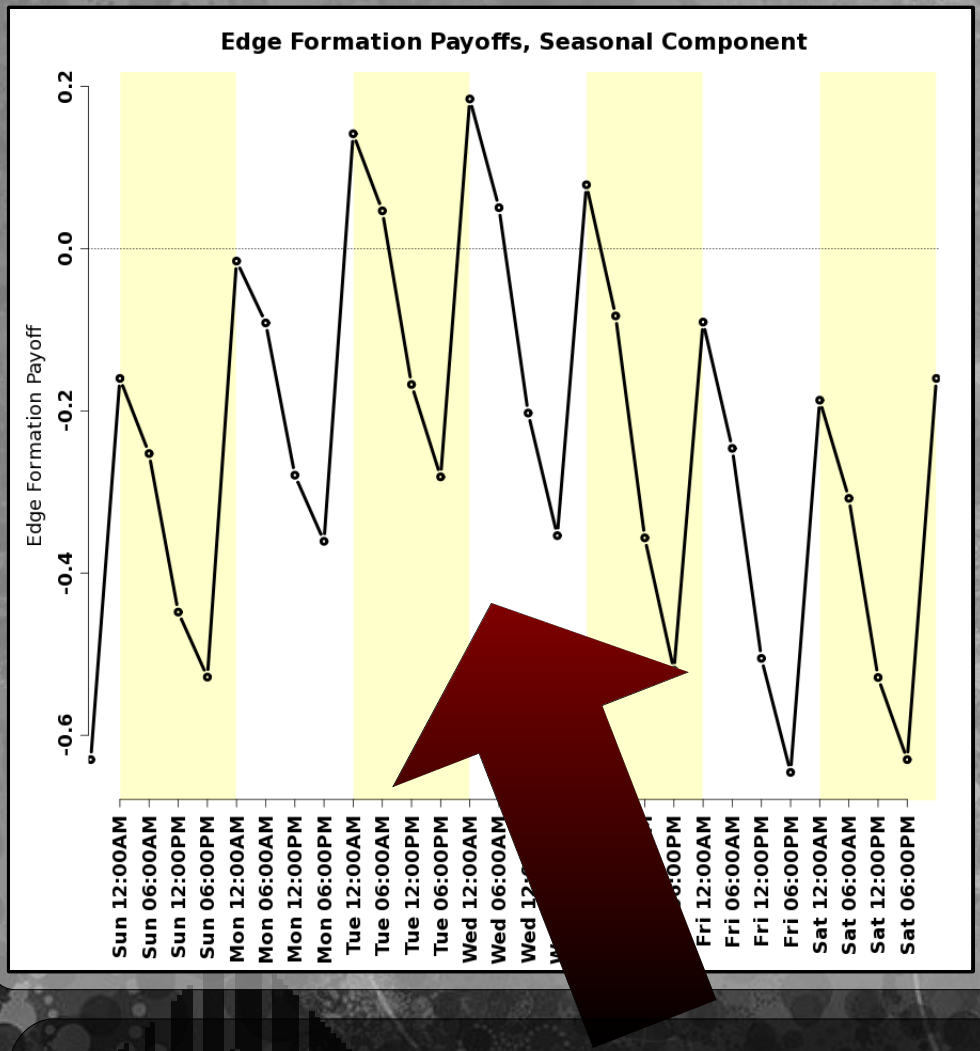
Temporal effects, including interactions and period effects

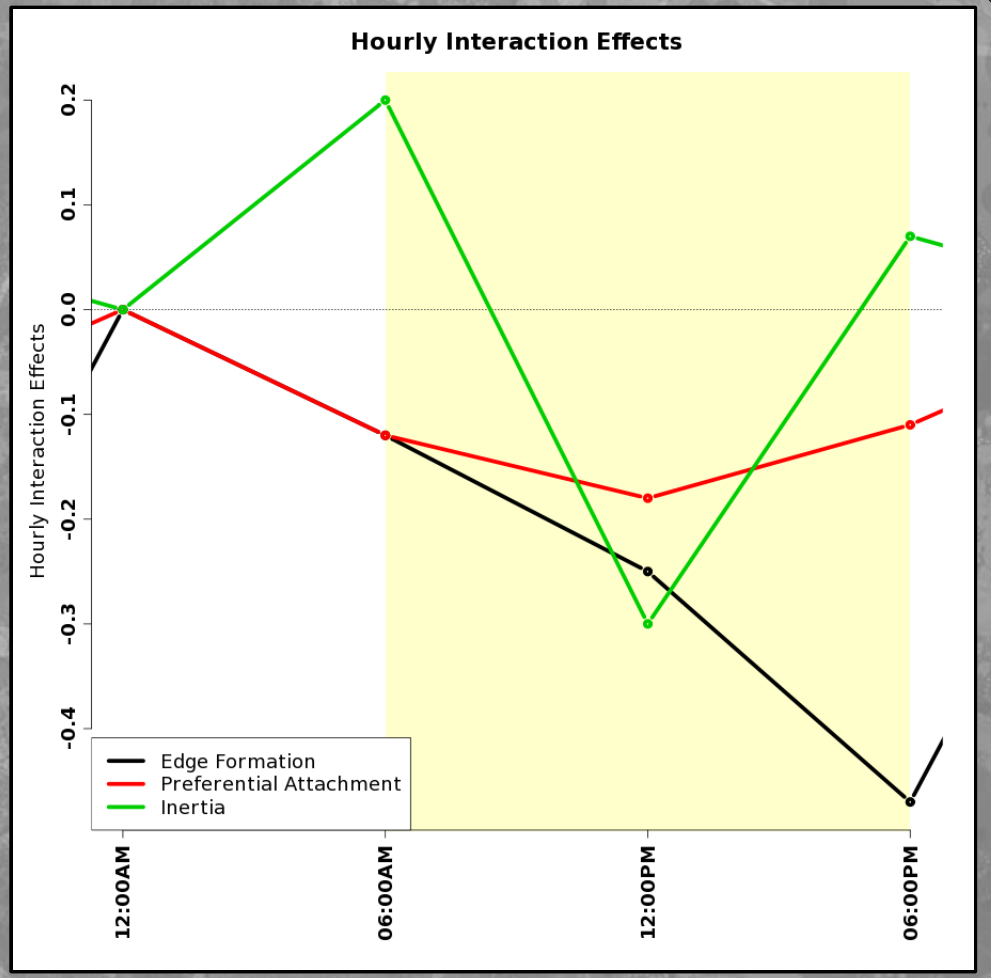
Behavioral Seasonality





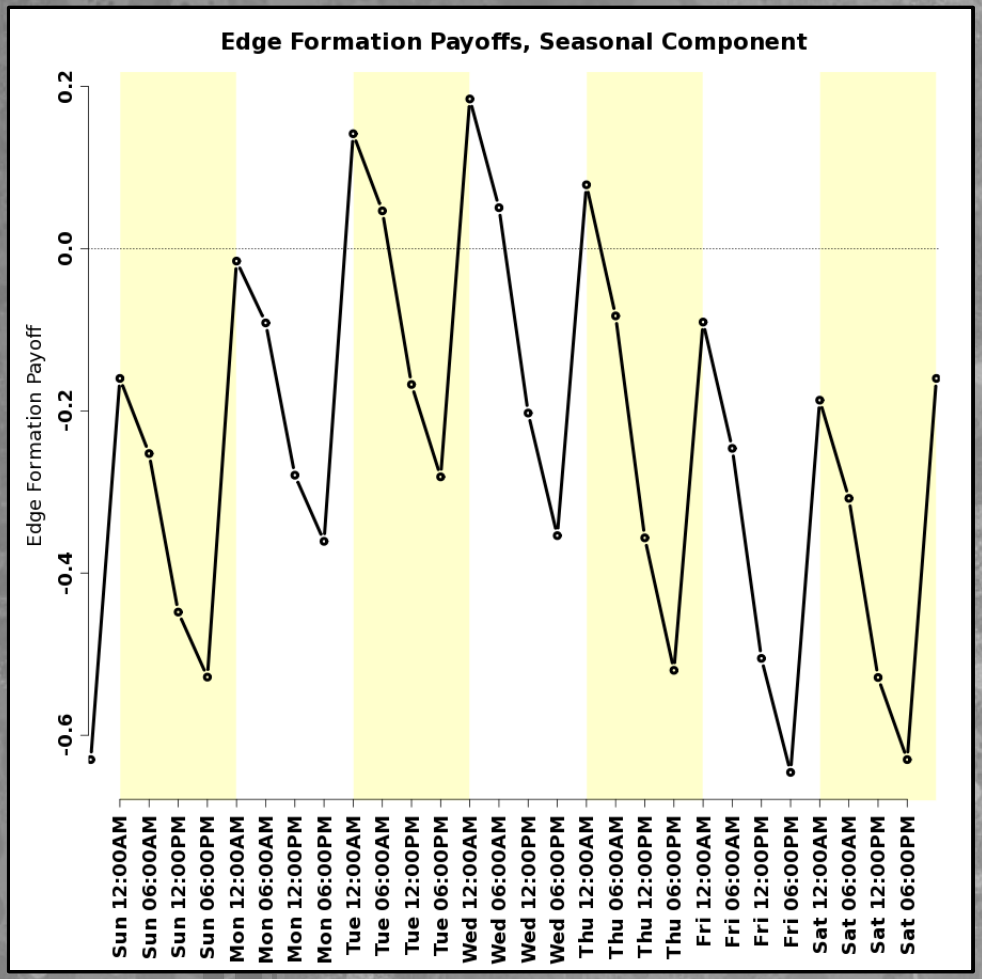
Behavioral Seasonality

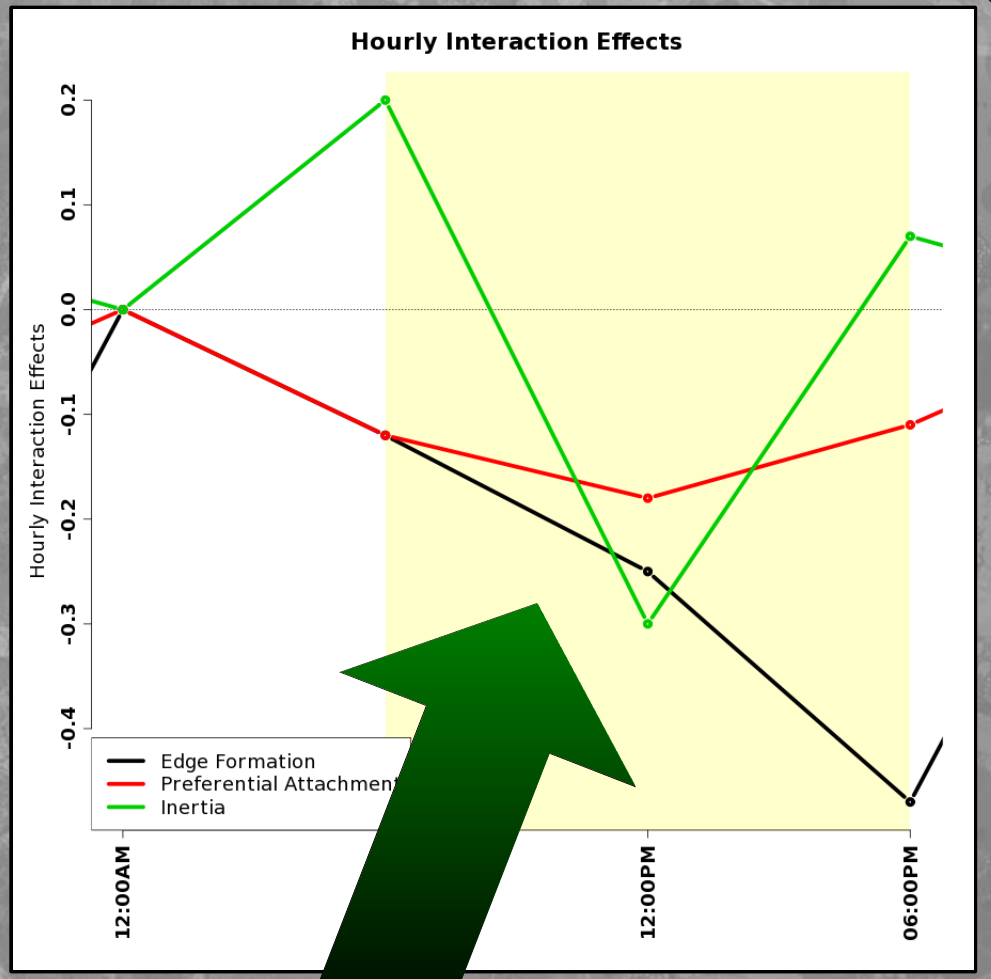




- Baseline propensity to form edges changes both by day and by week
- Highest at the start of the day, declining as the day goes on
- Builds and recedes during the week, lowest during the weekend

Behavioral Seasonality





- Other mechanisms also show daily seasonality
- Inertia highest in early/mid-morning, and in early evening
- Sensitivity to others' popularity highest at night (when tie formation tendencies also strongest); suggests attention shift

Directions and Discoveries: Vertex Dynamics

- Vertex set model critical for dynamic network prediction
  - Network size largely controls density, which shapes other factors
  - Vertex identities matter
    - Vertices carry their history and their covariates with them - errors in vertex prediction thus propagate into the edge model
    - When activity distributions skewed, accurate prediction of participation by a few "key players" can be critical to capturing other network properties
- Some current directions
  - Models for co-occurence (Smyth group)
  - "Open system" models from Dirichlet processes

Directions and Discoveries: Context Effects

- Internal dynamics are not enough
  - Need temporal, spatial, contextual covariates to predict networks in realistic settings
- Time matters in complex ways
  - Both edge structure and vertex set affected by past history
  - Mechanisms of action themselves change due to biological (e.g., sleep), institutional (e.g., work week), and other factors
- Geography matters, but not just as a baseline
  - In large networks, external forcing can be spatially contained; need to know when and where key events happen
- Some current directions
  - Dynamic latent feature models (Smyth group)
  - Geospatial computation, modeling for networks (Butts, Eppstein, Goodrich, Mount groups)

Conclusion

- We have made some exciting advances in our ability to model complex, dynamic networks
- Substantive advances as well as methodological ones
- Many new directions to pursue
- Problem illustrates benefits of our interdisciplinary approach