Some Comments on the Foundations of Network Analysis

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

Prepared for the August 25, 2009 UCI MURI AHM. This work was supported by DOD ONR award N00014-8-1-1015.

Background

- Many foundational issues in network analysis
  - Vertex, edge set definition, time scales, etc.
- Often taken for granted
  - Things "everyone knows" - but impact is not well-understood!
- Today: some comments from a recent review, and thoughts on how this affects our work

PERSPECTIVE

Revisiting the Foundations of Network Analysis

Carter T. Butts

Network analysis has emerged as a powerful way of studying phenomena as diverse as interpersonal interaction, connections among neurons, and the structure of the Internet. Appropriate use of network analysis depends, however, on choosing the right network representation for the problem at hand.

The past decade has seen a dramatic surge of interest in the study of networks, with much of it in fields outside the "traditional" areas of mathematics, computer science, and the social sciences (1, 2). By providing a formal mechanism for representation, measurement, and modeling of relational structure, the use of network analytic methods in these new domains (including physics, biology, and medicine) has arguably paved the way for a range of advances. On the other hand, this rapid expansion creates the risk that existing methods may be misapplied or misinterpreted, leading to inappropriate conclusions and generally poor results.

Standard Framework and Core Assumptions

Most network research is based on a representational formalism borrowed from graph theory. Researchers begin with a finite set of identifiable entities, which are represented via a vertex set. Each element of this set, commonly called a node, represents a single entity that potentially may take part in the relation under study. Relationships themselves are represented via edges, which conventionally are either unordered pairs of nodes (in which case the relation is said to be undirected) or ordered pairs of nodes (in which case the relation is said to be directed). The network is represented by a graph, which is defined as the set of nodes together with the set of pairwise relationships among them (Fig. 1A).

This representational framework is quite restrictive. To represent a system in this way, we must be able to reduce it to a well-defined set of discrete components whose interactions are strictly dyadic in nature. For any given (possibly ordered) pair of such components, the relationship is dichotomous, either present or absent; although such a framework may seem so restrictive as to be useless, its typical purpose is to serve as an approximation to the structure of a more complex system, for purposes of studying a particular property (such as the diffusion of a disease in a community over a specific time scale). Moreover, it is precisely the reductive nature of graphical structure that has facilitated its rich mathematical development (3) and associated scientific applications (4, 5).

Extensions and relaxations of this basic framework designed to accommodate more complex situations are many and varied. We may avoid the assumption of dichotomous relationships by allowing edges to carry different weights [such as the differing connection strengths among neurons in *Caenorhabditis elegans* (Fig. 1B) (6)]. Multilateral relationships (such as group memberships) may be represented by means of "hyperedges," which can involve arbitrarily many nodes (7). Temporal aspects of relationships may be handled by treating them as time series (8), such as with repeated cross-sectional sampling of group structure or Internet topology; as time intervals (9), such as with life history data on marital and employment relationships; or as effectively in… (10), such as with e-mail excha… munications (Fig. 1C).

Many measurement, analys… techniques are rooted within th… work. However, when assumpt… work do not serve as reasonab… of the system of interest, alter… tions and techniques may be ne… tors should be considered when c… representation, and what are the c… this choice is poorly made?

When Is a Node a Node?

Consider a biologist who wish… structure of animal parasite–pla… so undertakes a network study… time, technology, and resource… sample some designated area… such interactions, perhaps cons… of ties between each animal w… the sites on which it feeds. But… as a potential feeding site? Trea… a single site may seem reason… small plants but would obvious… tentially complex interactions as… a single tree. Additional detail… modated by distinguishing betw… tomical units (such as bark ve… leaves) or within classes, but… must be applied in determining… to make. The biologist's method of defining potential feeding sites will greatly influence the structure of the interaction network.

The basic problem is the definition of the class of distinct entities on which one's relation of interest will be defined. The mere act of positing such a class, of course, smuggles in the tacit assumption that such a class can be defined (and moreover, that it is scientifically useful to do so). The choice of individual humans as nodes in studies of friendship (11) or kinship (12) networks and the use of individual publications in citation studies (13) are examples in which this assumption is well-justified. On the other hand, studies of interactions between aggregates such as groups (14), households (15), or organizations may encounter problems due to the fluidity of the interacting units and the fact that subunits of a larger unit may themselves interact with others both within and without the "parent."

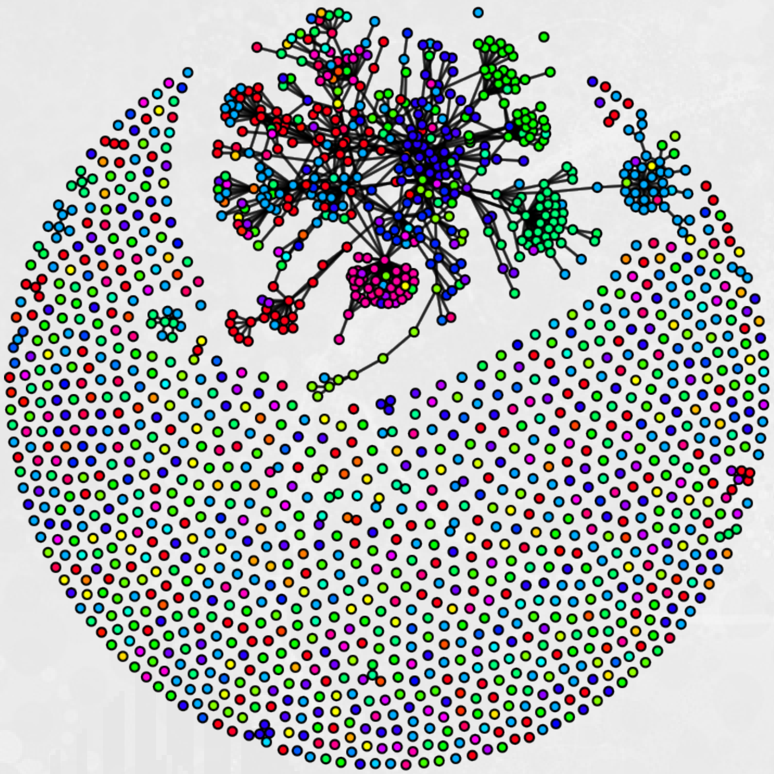
As in the biological example, collapsing all potentially interacting elements into a single unit may be a very poor approximation of reality. For example, my research group has studied networks formed during organizational responses to disasters. If we pooled all the groups operating under the aegis of one national government, then we would obscure the difference between small units such as urban search-and-rescue teams and large government ministries or departments, and also would incorrectly suggest that the resources or collaborators of one are necessarily available to

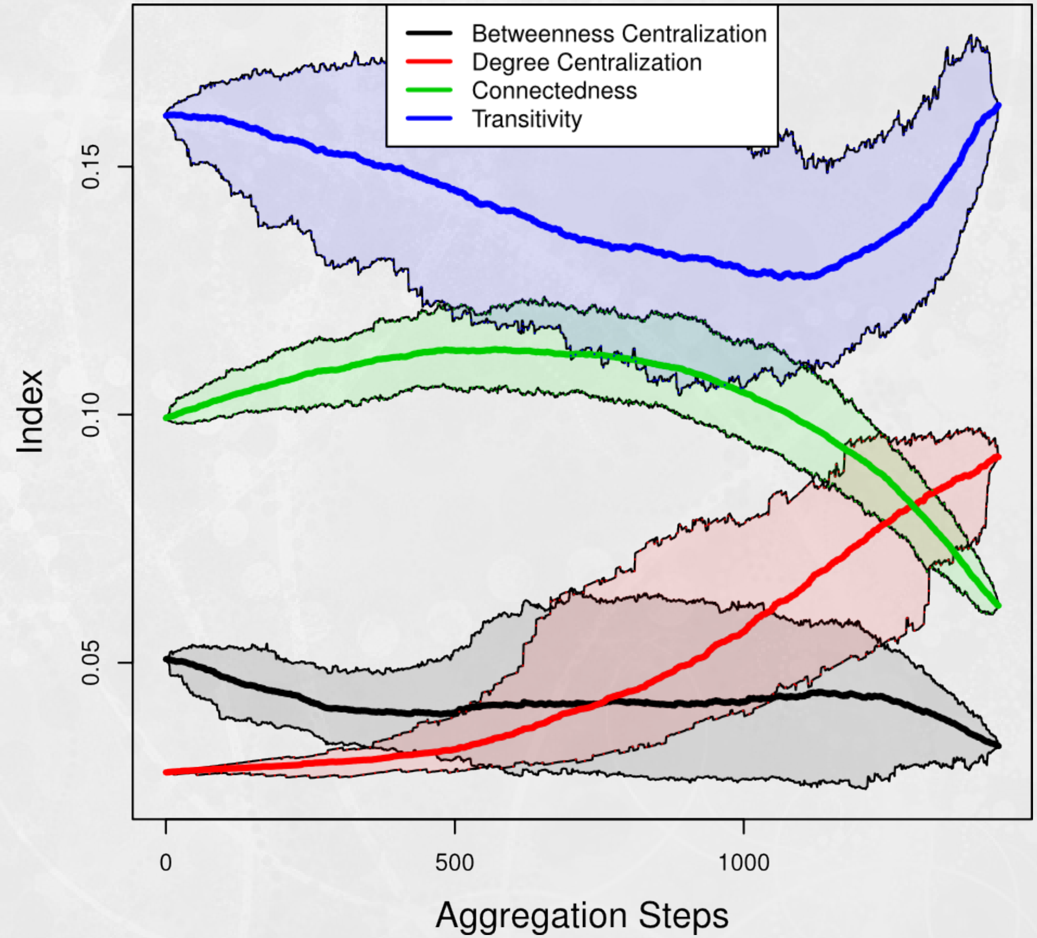
Department of Sociology and Institute for Mathematical Behavioral Sciences, University of California at Irvine, 3151 Social Science Plaza, Irvine, CA 92697–5100, USA. E-mail: buttsc@uci.edu

Choosing the Vertex Set

- Most basic issue - whence the vertex set?
- Not always obvious
  - Selection/boundary issues
  - Choice of scale in multiscale systems
  - Subordination among organizations
  - Containment/ recombination in households
- Different choices can greatly affect network measures

Effect of Vertex Aggregation

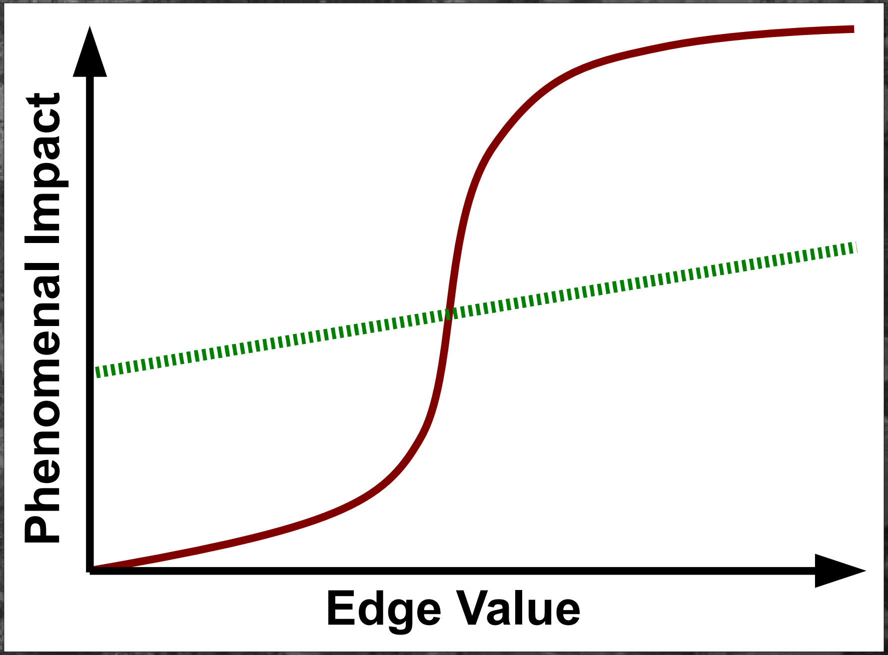




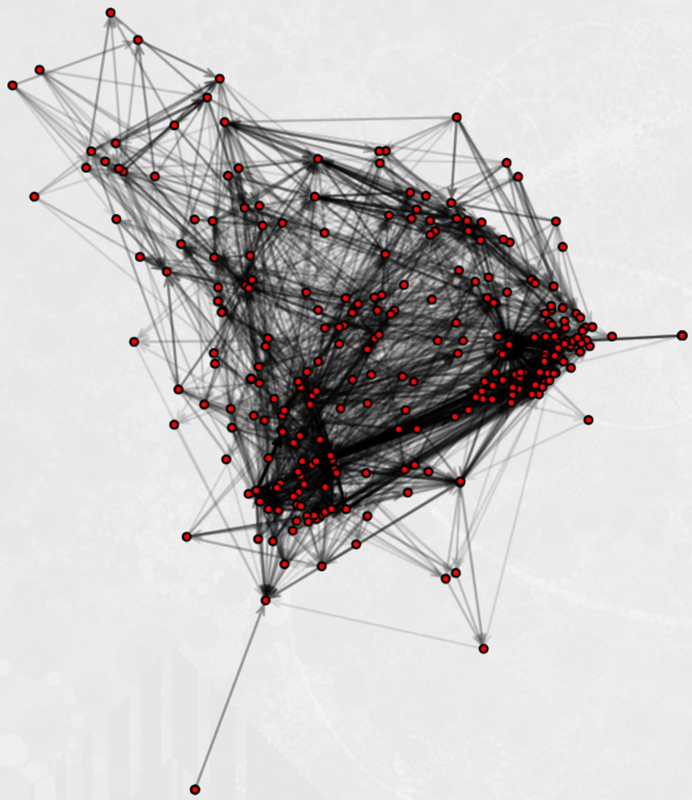
Katrina EMON Data (Butts et al., 2009)

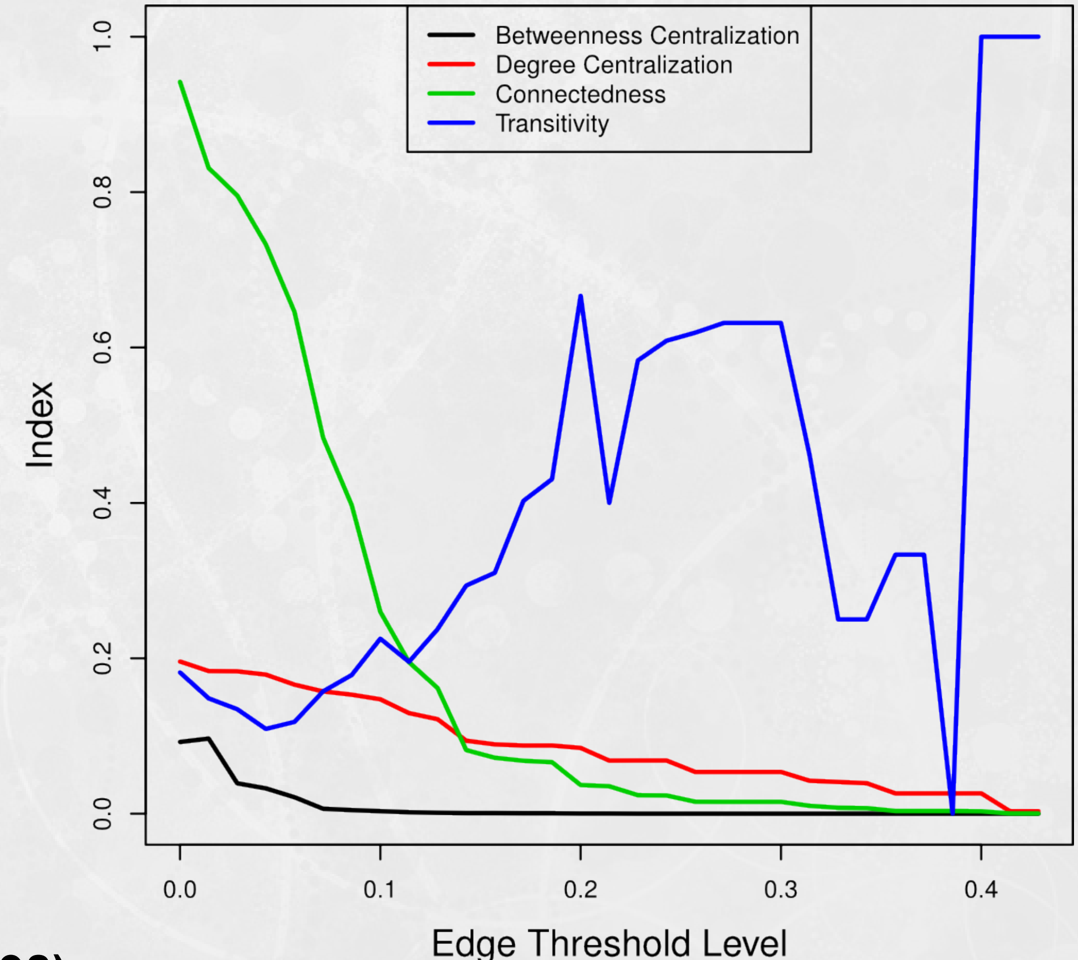
Valued Edges and Thresholding

- Well-recognized (but not well-understood) issue: dealing with valued edges
  - Most concepts, models dichotomous, but not all relations are
- Usual approach is thresholding, but this has non-obvious consequences....
  - Can be reasonable if edge behavior sigmoidal and threshold is well-chosen
  - Otherwise, same data can lead to completely different results



Effect of Threshold Selection





(*C. Elegans* Data (Watts, Strogatz, 1998)

Edge Timing and Network Processes

- Often, networks assumed as substrate for a social process
- May need to consider network dynamics...
  - Edge, vertex turnover
- ...but nature of dynamics depends on <u>relative</u> time scales of network, process evolution
  - Not whether network "is" static or dynamic in isolation
  - Right model can vary from fixed network to random mixing
    - Different models for different purposes

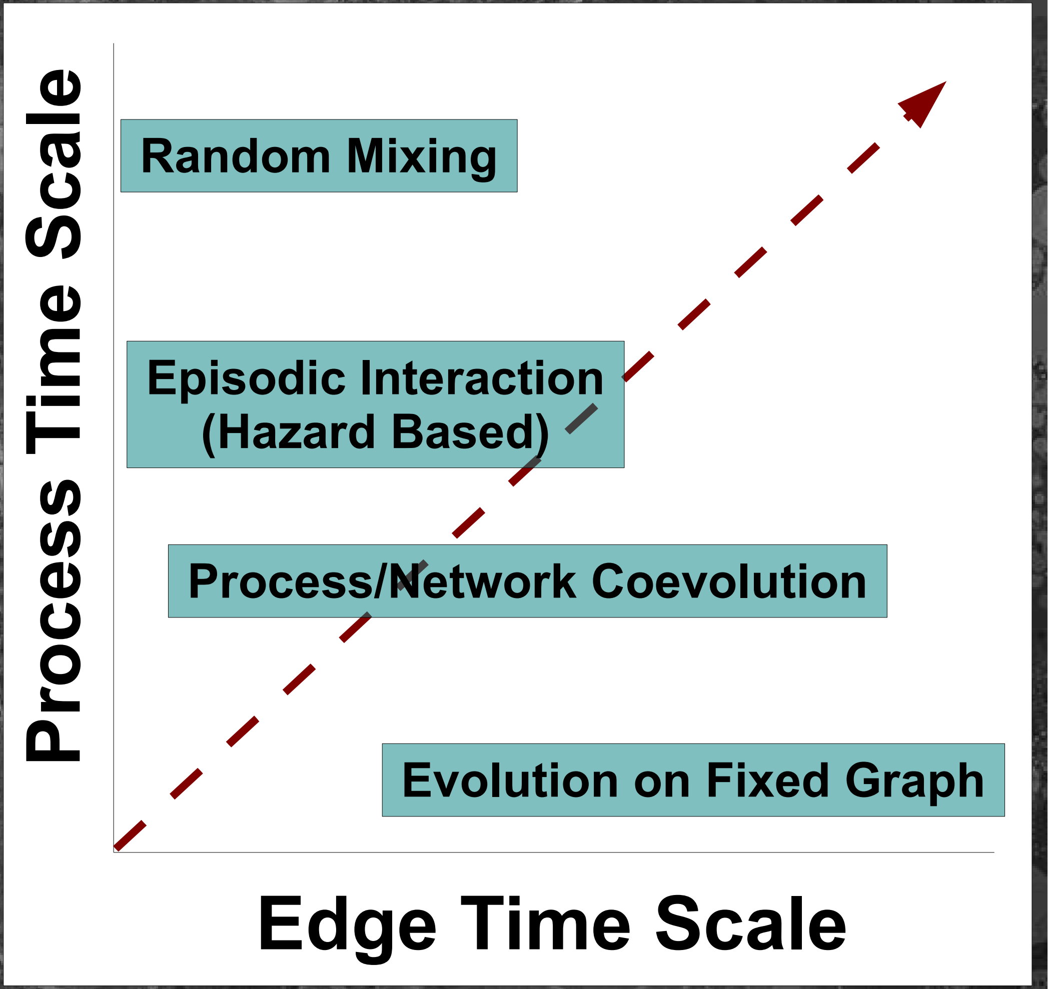
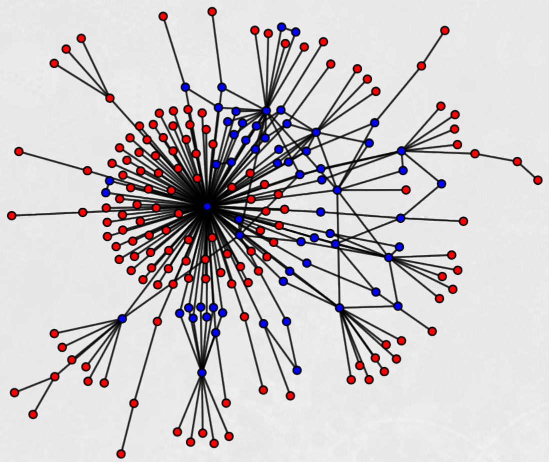


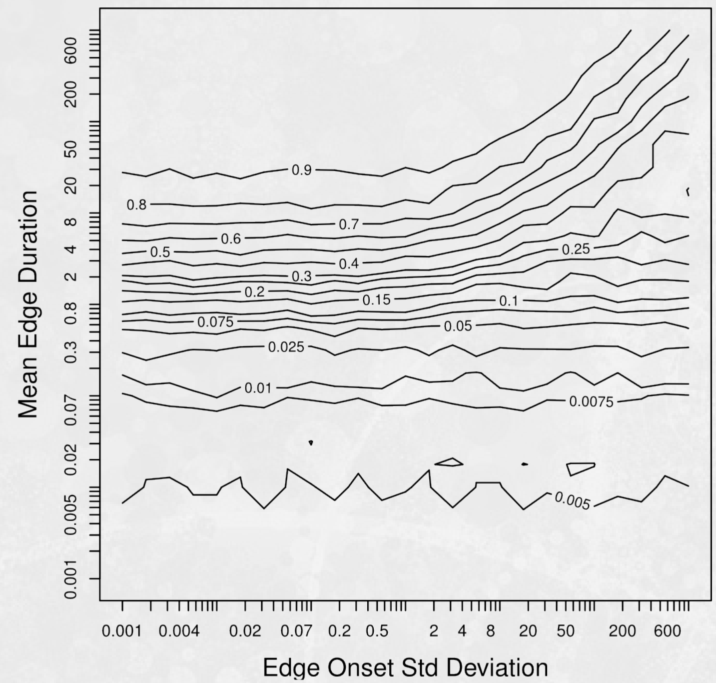
Illustration: Diffusion on an Evolving Network

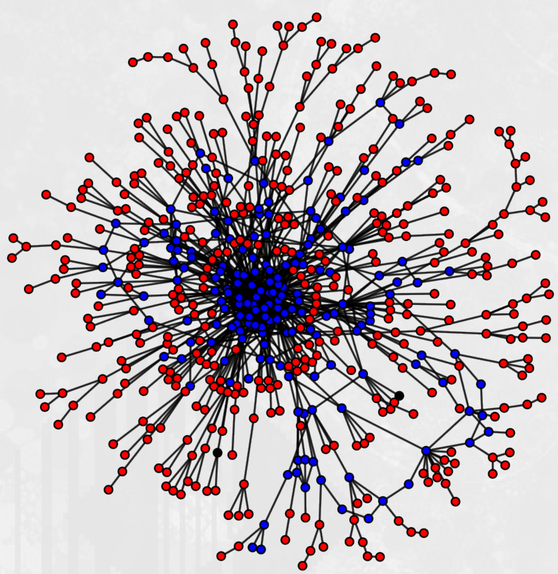
- Common process of interest: diffusion
  - Simple example:
    - Once "infected," vertices "infect" neighbors w/iid exponential waiting times
    - Process continues until all no available hosts left
  - Adding edge dynamics
    - Each relationship begins after iid exp waiting time, has iid exp duration
    - Intuition: edge dynamics affect permeability of network to diffusion

- How does edge timing affect diffusion?
  - Illustrative simulation:
    - Two sample networks
    - Mean duration, std dev of onset time varied
    - Poisson diffusion on dynamic network (starting at time 0) w/unit infection rate
  - Basic outcome: expected fraction of population infected by a single, random "seed"
    - How powerful are timing effects?

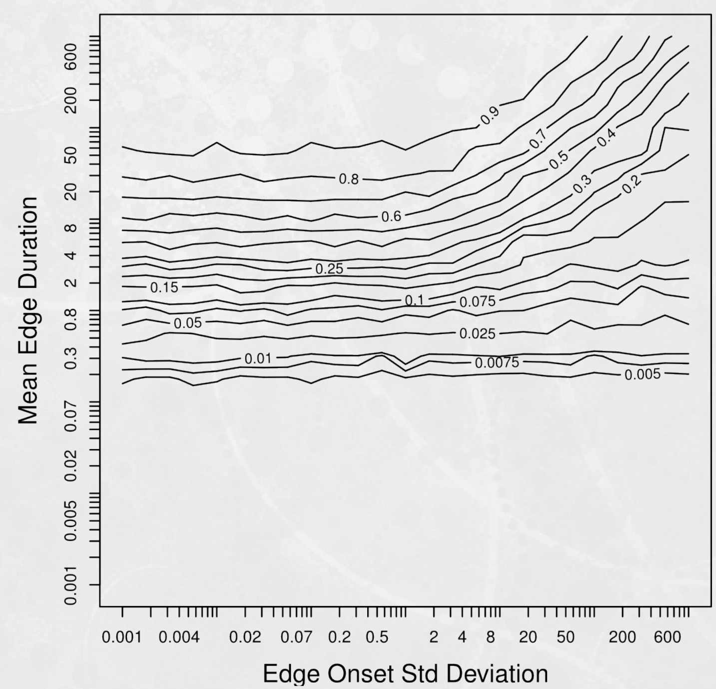


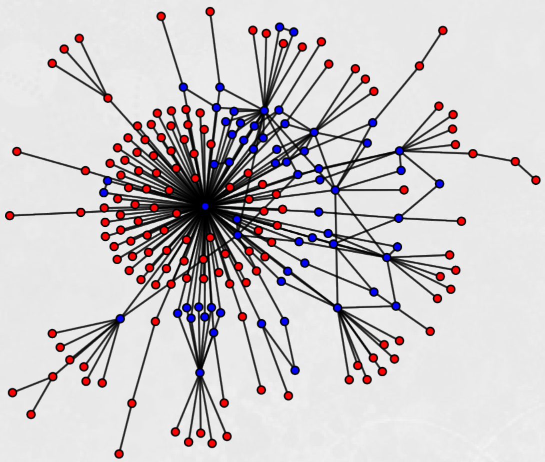
WTC Radio Data
(Butts et al., 2007)



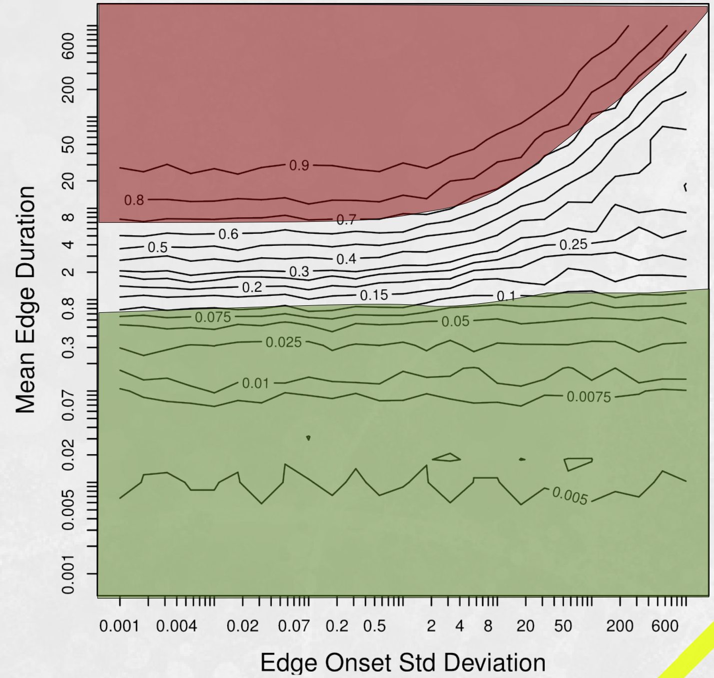


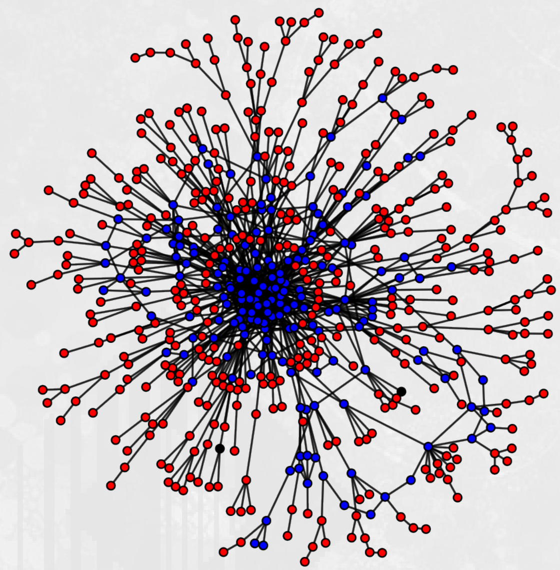
Sexual Contact Data
(Potterat et al., 2002)



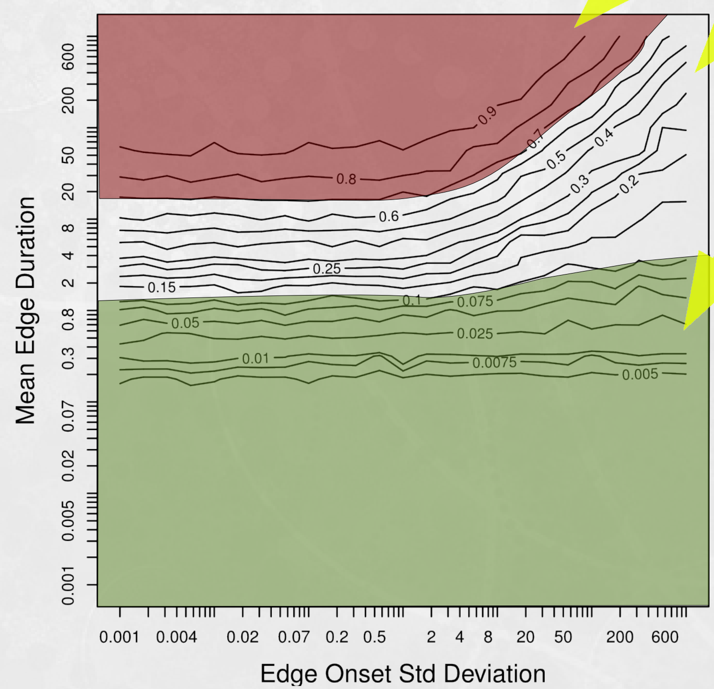


WTC Radio Data
(Butts et al., 2007)





Sexual Contact Data
(Potterat et al., 2002)



- Time scales determine diffusion behavior
- Three basic regimes
  - Near-complete diffusion
  - Incomplete diffusion
  - Minimal diffusion
- Behavior similar across networks
  - Differences in degree distribution, clustering, cohesion matter less than timing!

Some Conclusions and Project-Related Comments

- General conclusions:
  - Need to be attentive to the basics
    - "Any old network" may be OK for algorithm testing, but not for serious analysis
  - Need to learn more about robustness of methods to "network specification error"
  - May need models for alternate data representations

- Project-specific recommendations:
  - Simple models for valued data?
    - "Threshold regression" ERGMs?
  - Models for vertices w/containment or hierarchical structure
    - Not sure that block-hierarchical ERGMs enough, but a start
  - Keep pushing on dynamics!