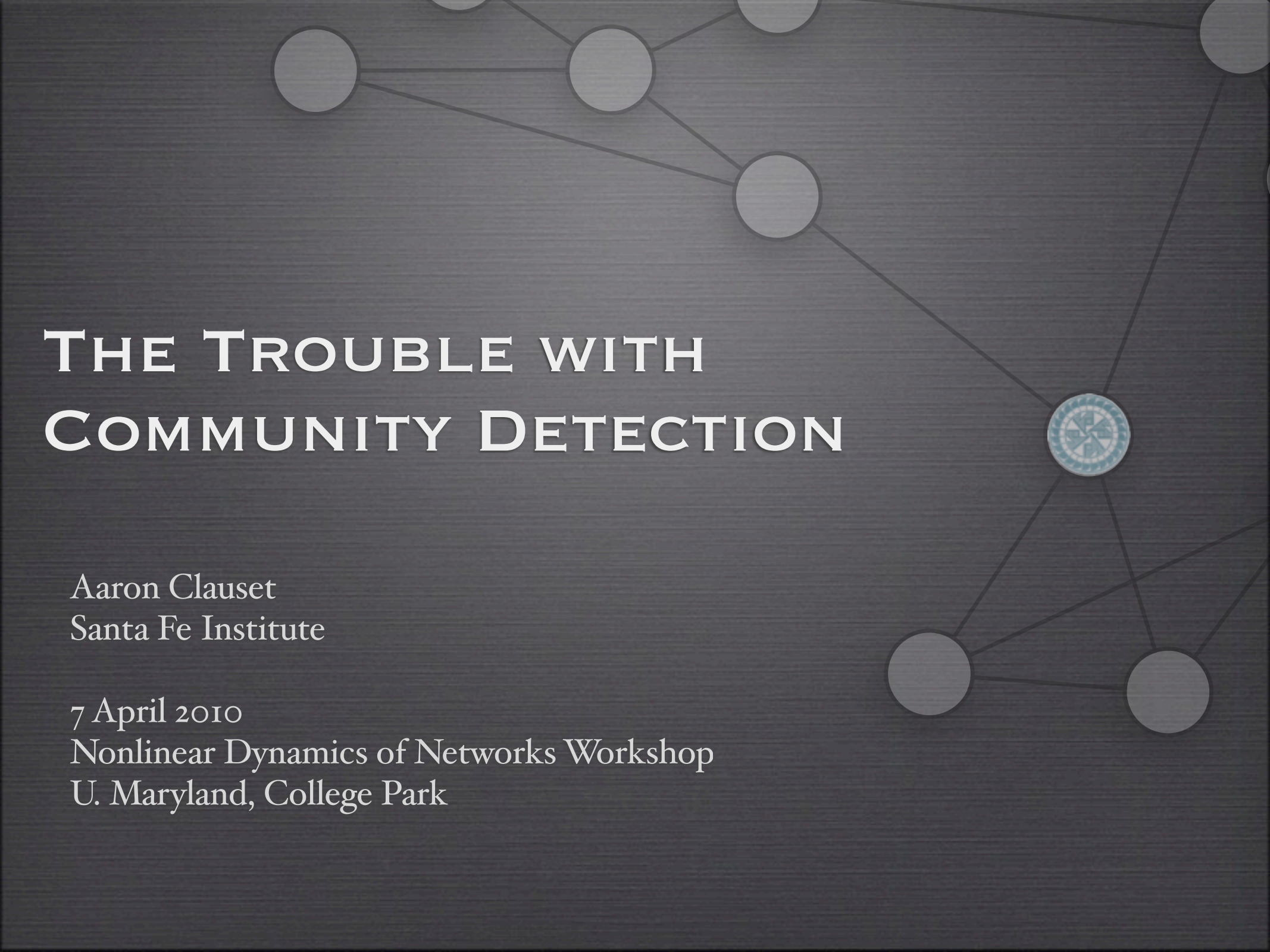


The Trouble with Community Detection

Aaron Clauset
Santa Fe Institute

7 April 2010
Nonlinear Dynamics of Networks Workshop
U. Maryland, College Park

Thanks to

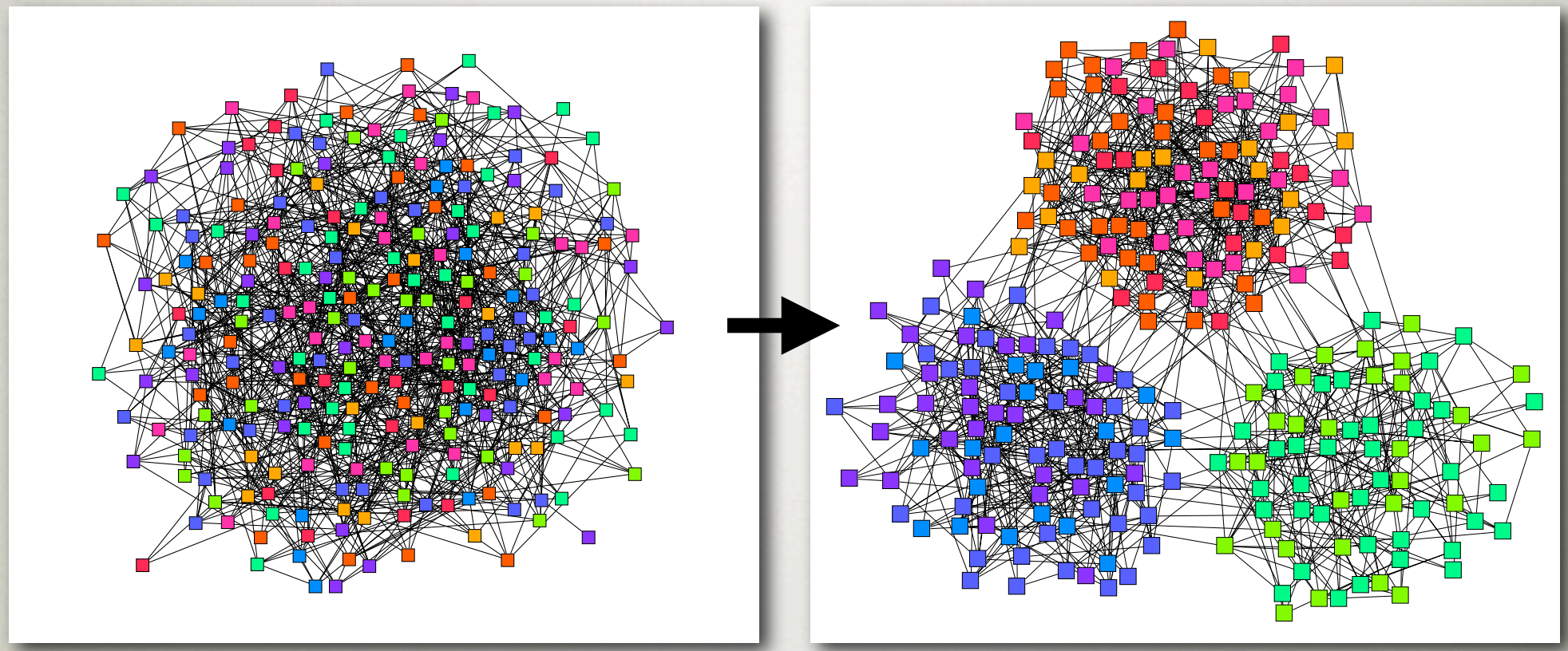
- National Science Foundation REU Program
- James S. McDonnell Foundation
- Vincent Blondel
- Joint work with

Ben Good (Swarthmore)

Yva de Montjoye (MIT)

"The performance of modularity maximization in practical contexts."
B.H. Good, Y.-A. de Montjoye, A. Clauset, *Phys. Rev. E*, to appear.
arxiv:0910.0165

Community/Module/Compartment Identification



- social web communities
- semantic groupings
- predicting node labels
- functional modules (biology)
- etc.

Newman-Girvan modularity

$$Q = \sum_{i=1}^{k} \left[\frac{e_i}{m} - \left(\frac{d_i}{2m} \right)^2 \right]$$

- observed within-module density vs. expected within-module density
- degree-preserving random graph as null model
- finding "good" partition = optimize Q over partitions
- *de facto* standard (1431+ Google Scholar citations; n.b., many other techniques exist)

Newman and Girvan 2004

Newman-Girvan modularity

$$Q = \sum_{i=1}^{k} \left[\frac{e_i}{m} - \left(\frac{d_i}{2m} \right)^2 \right]$$

- NP-hard, but many heuristics work well in practice:
  - greedy agglomeration
  - mathematical programming
  - spectral optimization
  - extremal optimization
  - simulated annealing
  - sampling (MCMC, etc.)
  - ...

Brandes et al. 2008, Newman 2004, Clauset et al. 2004, Blondel et al. 2008, Agarwal and Kempe 2008, Newman 2006, Richardson et al. 2009, Duch and Arenas 2005, Guimera and Amaral 2005, Massen and Doye 2006, Sales-Pardo et al. 2007

Newman-Girvan modularity

$$Q = \sum_{i=1}^{k} \left[\frac{e_i}{m} - \left(\frac{d_i}{2m} \right)^2 \right]$$

- In practice, three common assumptions:

  1. global maximum is the "best" partition
  2. modular networks have clear, global maximum
  3. high-modularity partitions structurally similar

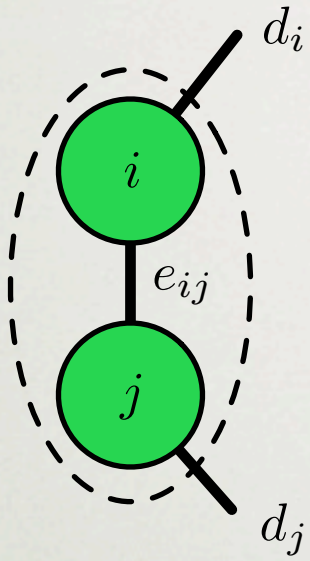
- All three are wrong.

This talk:

1. Is the maximum the “best” partition?
2. A clear global maximum?
3. How many high-modularity partitions?
4. Is the modularity function smooth?
5. How similar are high-modularity partitions?

1. Is the maximum the "best" partition?

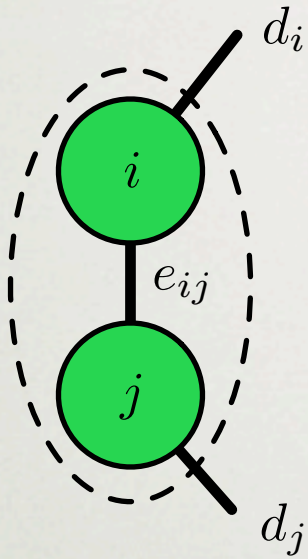
Consider merging two strong modules (e.g., cliques)



$$\Delta Q_{ij} = \frac{e_{ij}}{m} - \frac{d_i d_j}{2m^2}$$

1. Is the maximum the "best" partition?

Consider merging two strong modules (e.g., cliques)



$$\Delta Q_{ij} = \frac{e_{ij}}{m} - \frac{d_i d_j}{2m^2}$$

Merging always favored when

$$e_{ij} > \frac{d_i d_j}{2m} \qquad \text{i.e.,} \qquad e_{ij} > \mathrm{E}[e_{ij}]$$

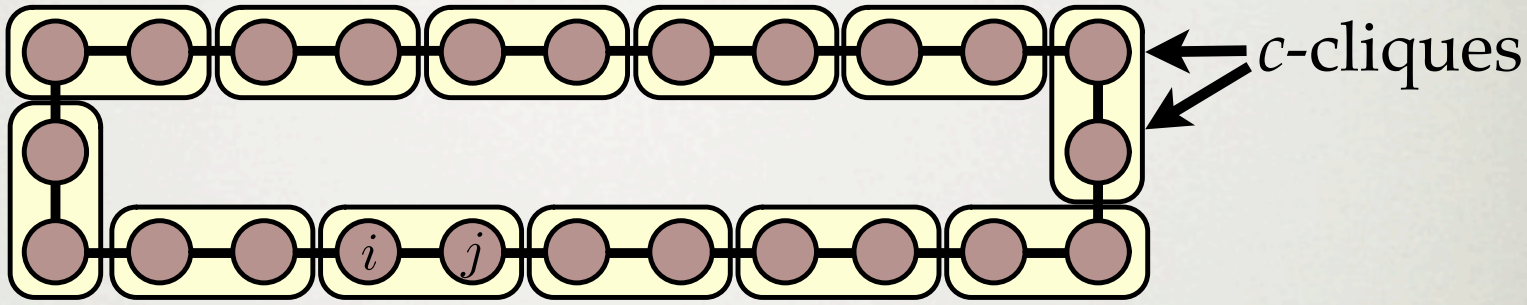
Thus, $Q_{\max}$ ***won't distinguish modules*** i ***and*** j ***and*** Q ***exhibits a "resolution limit"***

Fortunato and Barthelemy 2007, Kumpula et al. 2007, Branting 2008, Berry et al. 2009

1. Two examples of the resolution limit

1. A first example

A "ring" of k cliques of c nodes each

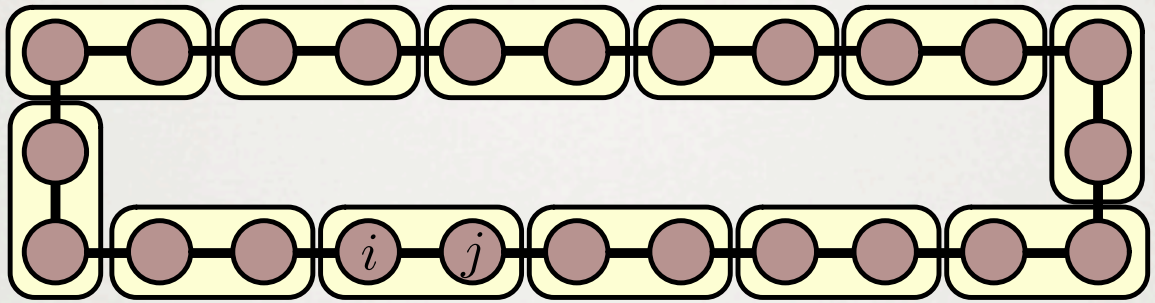


where $e_{ij} = 1$ and $e_i = \binom{c}{2}$

Thus $d_i = \binom{c}{2} + 2$

1. A first example

A "ring" of k cliques of c nodes each



When merging adjacent cliques:

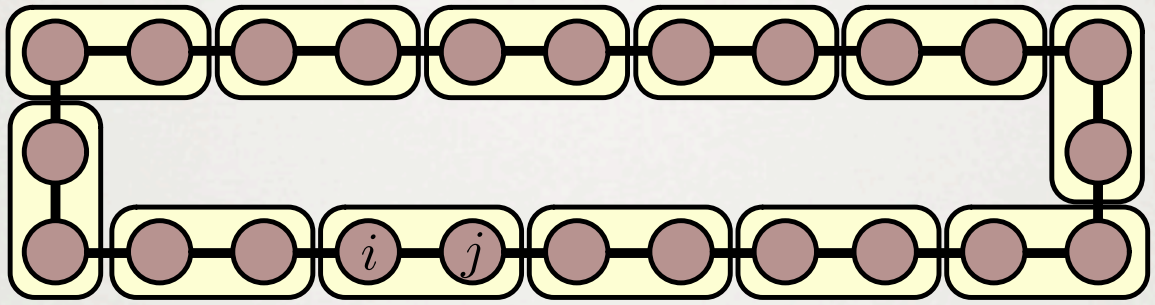
$$\Delta Q = \frac{1}{k\left[\binom{c}{2}+1\right]} - 2k^{-2}$$

$\Delta Q > 0$ whenever

$$k > 2\binom{c}{2} + 2$$

1. A first example

A "ring" of k cliques of c nodes each



When merging adjacent cliques:

$$\Delta Q = \frac{1}{k\left[\binom{c}{2}+1\right]} - 2k^{-2}$$

$\Delta Q > 0$ whenever

$$k > 2\binom{c}{2} + 2$$

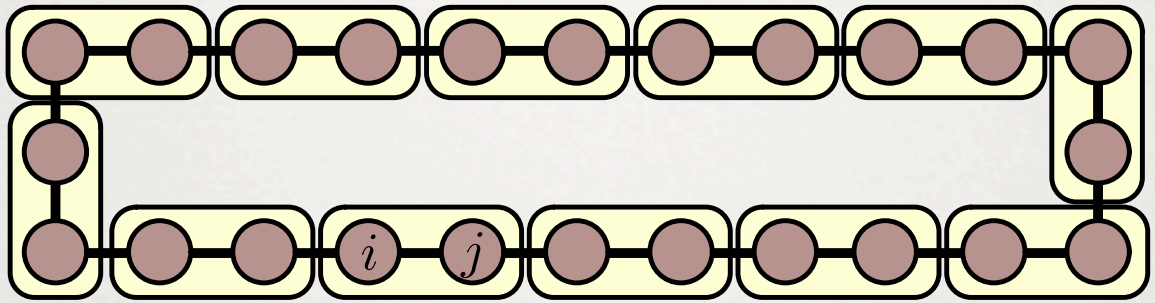
For $k = 24$ and $c = 5$:

Each clique in a group, $Q_1 = 0.8674$

Pairs of cliques together, $Q_2 = 0.8712$

1. A second example

A "ring" of k cliques of c nodes each

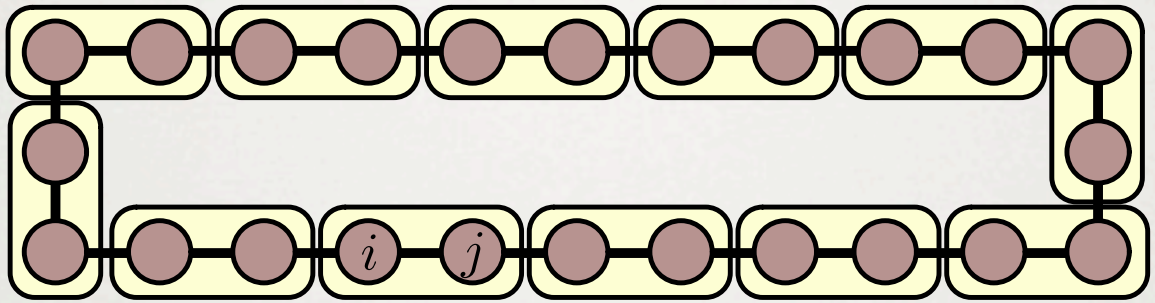


where $e_{ij} = 2/(k-1)$ and $e_i = \binom{c}{2}$

Thus $d_i = \binom{c}{2} + 2$

1. A second example

A "ring" of k cliques of c nodes each



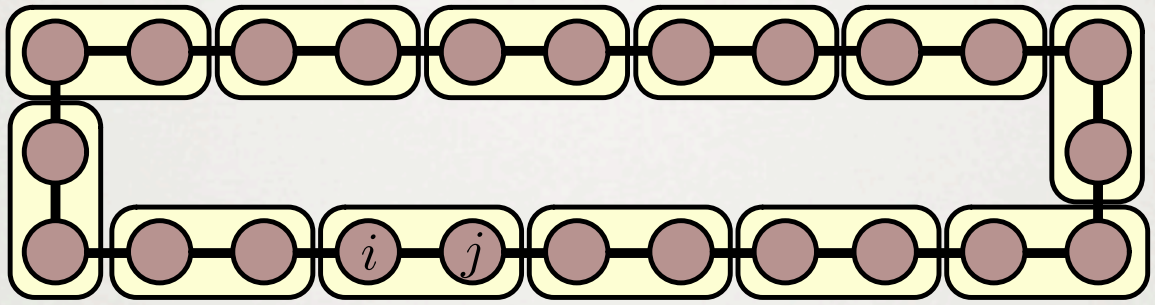
When merging adjacent cliques:

$$\Delta Q = \frac{2}{k(k-1)\left[\binom{c}{2}+1\right]} - 2k^{-2} \qquad \Delta Q > 0 \text{ only for } k \leq 2$$

Thus, **no resolution limit**. Why?

1. A second example

A "ring" of k cliques of c nodes each



When merging adjacent cliques:

$$\Delta Q = \frac{2}{k(k-1)\left[\binom{c}{2}+1\right]} - 2k^{-2} \qquad \Delta Q > 0 \text{ only for } k \leq 2$$

Thus, **no resolution limit**. Why?

Recall: $e_{ij} = 2/(k-1)$ and $\mathrm{E}[e_{ij}] = O(k^{-1})$

1. Take home messages

1. Resolution limit not universal
   - Appears mainly in large, unweighted networks

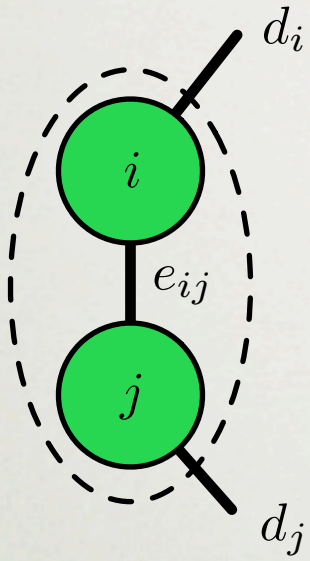
     where $e_{ij} = O(1)$ but $\mathrm{E}[e_{ij}] = O(k^{-1})$
2. Q measures deviations from random graph model

$$Q = \sum_{i=1}^{k} \left[\frac{e_i}{m} - \left(\frac{d_i}{2m} \right)^2 \right]$$

3. Res. limit = clash between intuition and definition
4. Appears in most other objective functions (e.g., Potts models, likelihood functions, etc.)
   - Need alternatives; local methods?

2. A clear, global maximum?

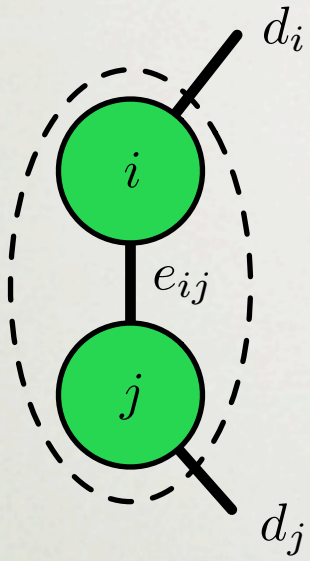
Consider merging two strong modules



$$\Delta Q_{ij} = \frac{e_{ij}}{m} - \frac{d_i d_j}{2m^2}$$

2. A clear, global maximum?

Consider merging two strong modules



$$\Delta Q_{ij} = \frac{e_{ij}}{m} - \frac{d_i d_j}{2m^2}$$

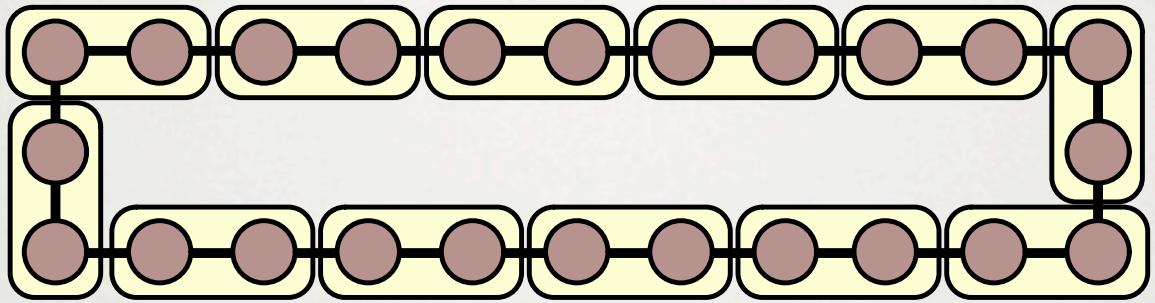
For roughly balanced groups $d_i \approx 2m/k$, and

$$\Delta Q_{ij} \geq -2k^{-2}$$

Thus, these partitions have $Q \approx Q_{\max}$

2. An example

A "ring" of *k* cliques of *c* nodes each



For $k = 24$ and $c = 5$:

Each clique in a group: $Q_1 = 0.8674$

Pairs of cliques together: $Q_2 = 0.8712$

$$\Delta Q = -0.0038$$

3. How many high-modularity partitions?

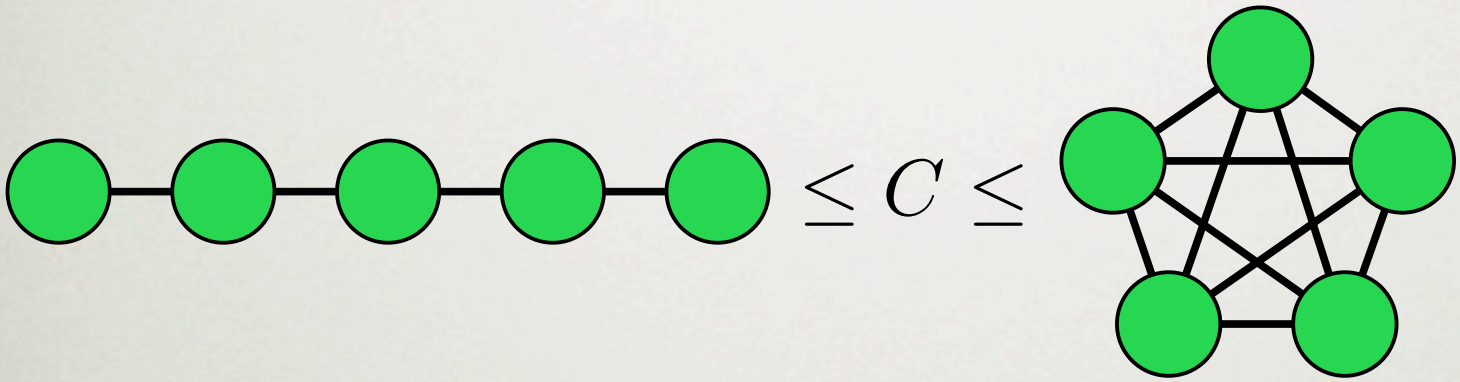
3. How many high-modularity partitions?

$$2^{k-1} \leq C$$

$\leq C$

3. How many high-modularity partitions?

$$2^{k-1} \leq C \leq B_k \text{ (}k\text{th Bell number)}$$



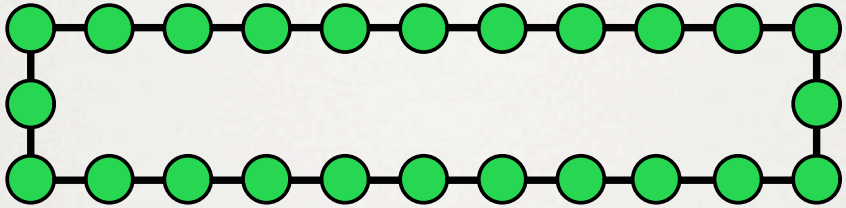
Exponentially many (or more) high-modularity (degenerate) solutions.

4. Is the modularity function smooth?

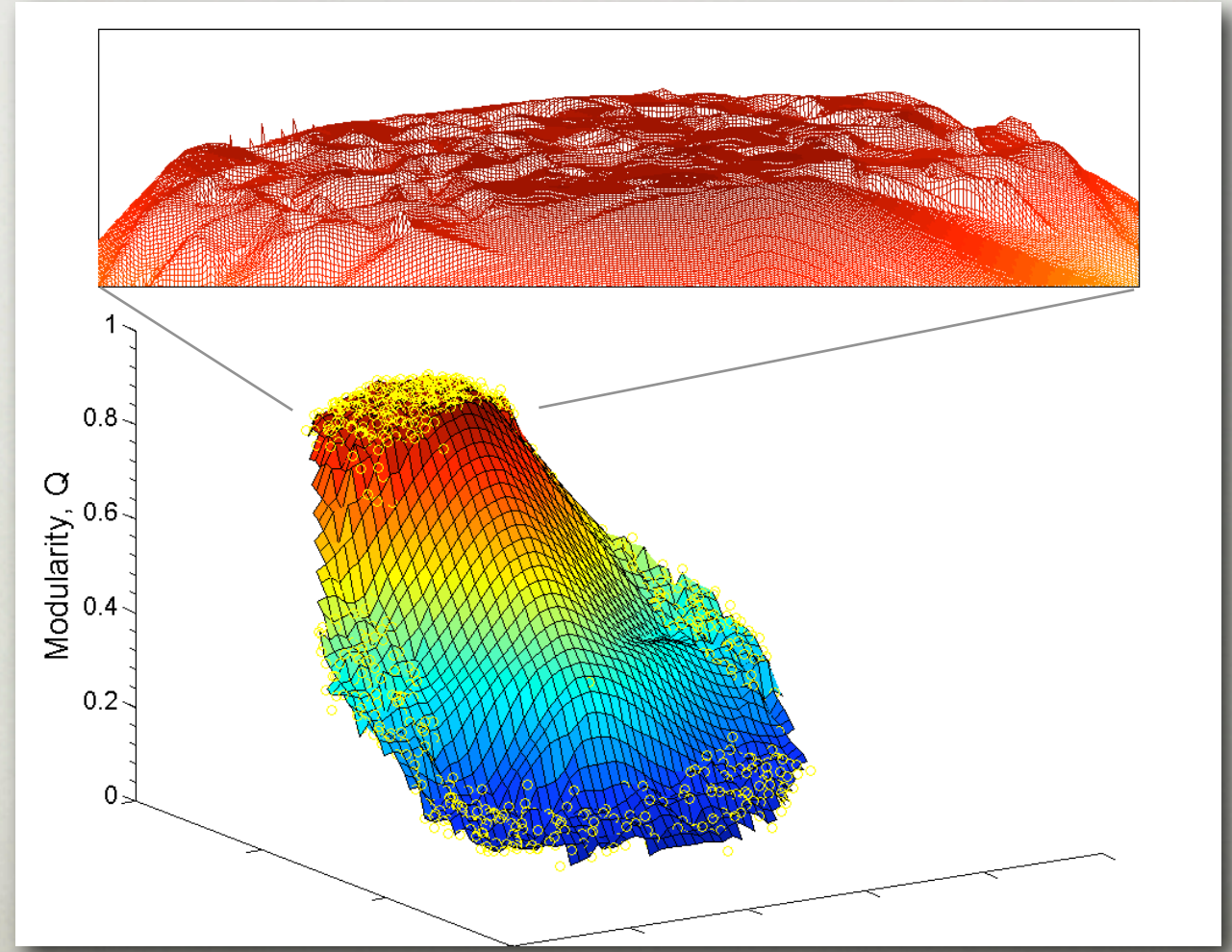
4. Is the modularity function smooth?

1. choose a network
2. sample many high modularity partitions
   (e.g., via simulated annealing)
3. embed them in 2D Euclidean space,
   such that pairwise distances are preserved
   (e.g., via curvilinear component analysis)

4. Is the modularity function smooth?

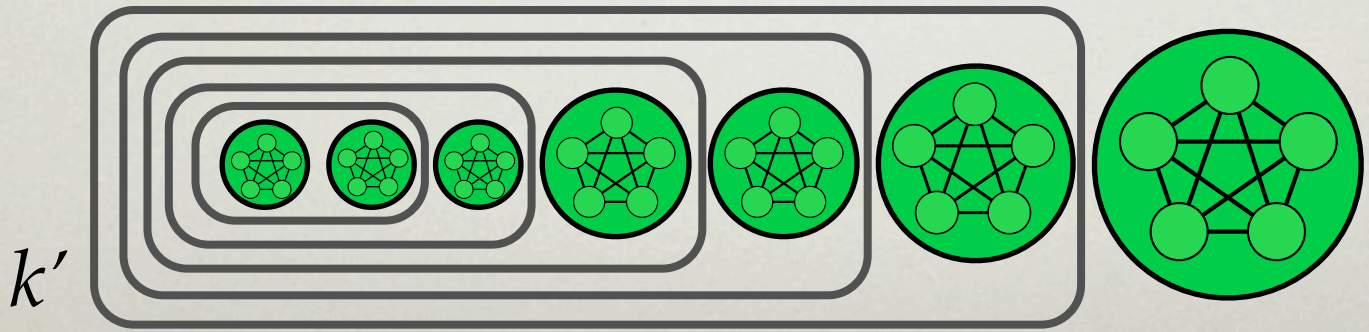


("ring" network)

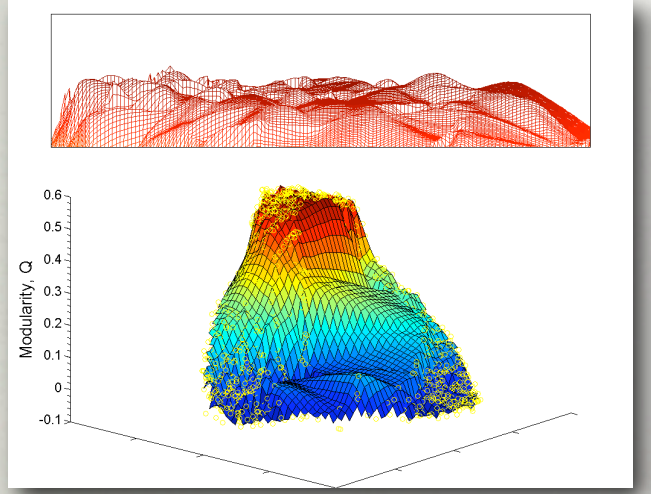


5. How similar are high-modularity partitions?

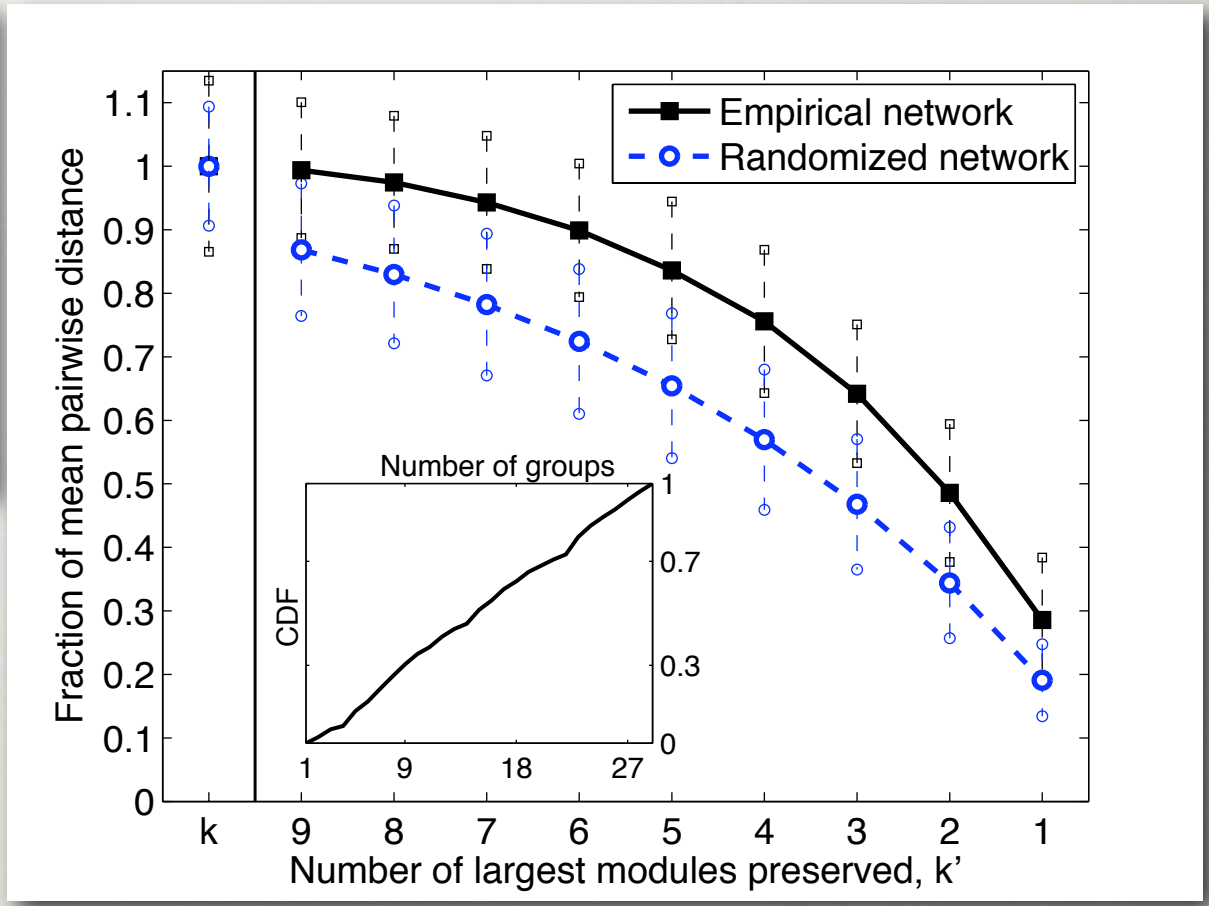
1. choose a real-world network
   (metabolic network for spirochaete *T. pallidum)*
2. sample many high modularity partitions
   (via simulated annealing)
3. for each partition, merge all but k' largest modules; compute mean pairwise distance as function of k': $\langle d(C_1, C_2) \rangle_{k'}$



5. How similar are high-modularity partitions?



for *T. pallidum*



Conclusions:

Conclusions:

1. modularity function is highly degenerate
    - degeneracies appear in other score functions, including local methods and generative models

Good, de Montjoye and Clauset "The performance of modularity maximization in practical contexts." *Physical Review E* (2010). **arxiv:0910.0165**

Conclusions:

1. modularity function is highly degenerate
    - degeneracies appear in other score functions, including local methods and generative models
2. more modular = more degenerate
    - degeneracies are caused by choice of null model

Good, de Montjoye and Clauset "The performance of modularity maximization in practical contexts." *Physical Review E* (2010). **arxiv:0910.0165**

Conclusions:

1. modularity function is highly degenerate
    - degeneracies appear in other score functions, including local methods and generative models
2. more modular = more degenerate
    - degeneracies are caused by choice of null model
3. good partitions easy to find
    - exponential in number

Good, de Montjoye and Clauset "The performance of modularity maximization in practical contexts." *Physical Review E* (2010). **arxiv:0910.0165**

Conclusions:

1. modularity function is highly degenerate
    - degeneracies appear in other score functions, including local methods and generative models
2. more modular = more degenerate
    - degeneracies are caused by choice of null model
3. good partitions easy to find
    - exponential in number
4. but they're structurally different
    - any one partition should not be trusted

Good, de Montjoye and Clauset "The performance of modularity maximization in practical contexts." *Physical Review E* (2010). **arxiv:0910.0165**

Conclusions:

1. modularity function is highly degenerate
    - degeneracies appear in other score functions, including local methods and generative models
2. more modular = more degenerate
    - degeneracies are caused by choice of null model
3. good partitions easy to find
    - exponential in number
4. but they're structurally different
    - any one partition should not be trusted
5. thus, optimization alone is misleading
    - we need to combine information across many solutions.
6. we need new approaches

Good, de Montjoye and Clauset "The performance of modularity maximization in practical contexts." *Physical Review E* (2010). **arxiv:0910.0165**