Can spin glass technology help deep learners find small communities?

Scott Kirkpatrick
Hebrew University of Jerusalem

Cargese, Sept. 2014

Communities exist on all scales

- Consider Europe, North America -- 300M people, divides into:
  - Those who might vote liberal, or conservative
  - Who might buy a model of car
  - Who might be vulnerable to a common disease… all $\propto N$
- But suppose you are looking for
  - People who lease private jets? 15-30,000 people in this market
  - People vulnerable to a rare but genetically linked disease?
  - These groups may be $\propto \sqrt{N}$ but probably still unique.
- Finally, consider things on the scale of $log_2 N$
  - 30 people.
  - Generally we search from bottom up, and must focus attention on one cluster which is not unique.
  - E.g. Terrorist sleeper cells -- you find one bad guy and look for his really close friends.
  - Can message-passing accelerate this search?

Replicas and Cavity Approach?

- These separate small differences in large complex systems. Separate $\sqrt{N}$ signal from $\sqrt{N}$ noise. Do they also single out small things?

Spin Glass Theory and Beyond (World Scientific Lecture Notes in Physics) by Marc Mezard, Giorgio Parisi and Miguel Angel Virasoro (Nov 1987)

$1,996.01 used (3 offers)

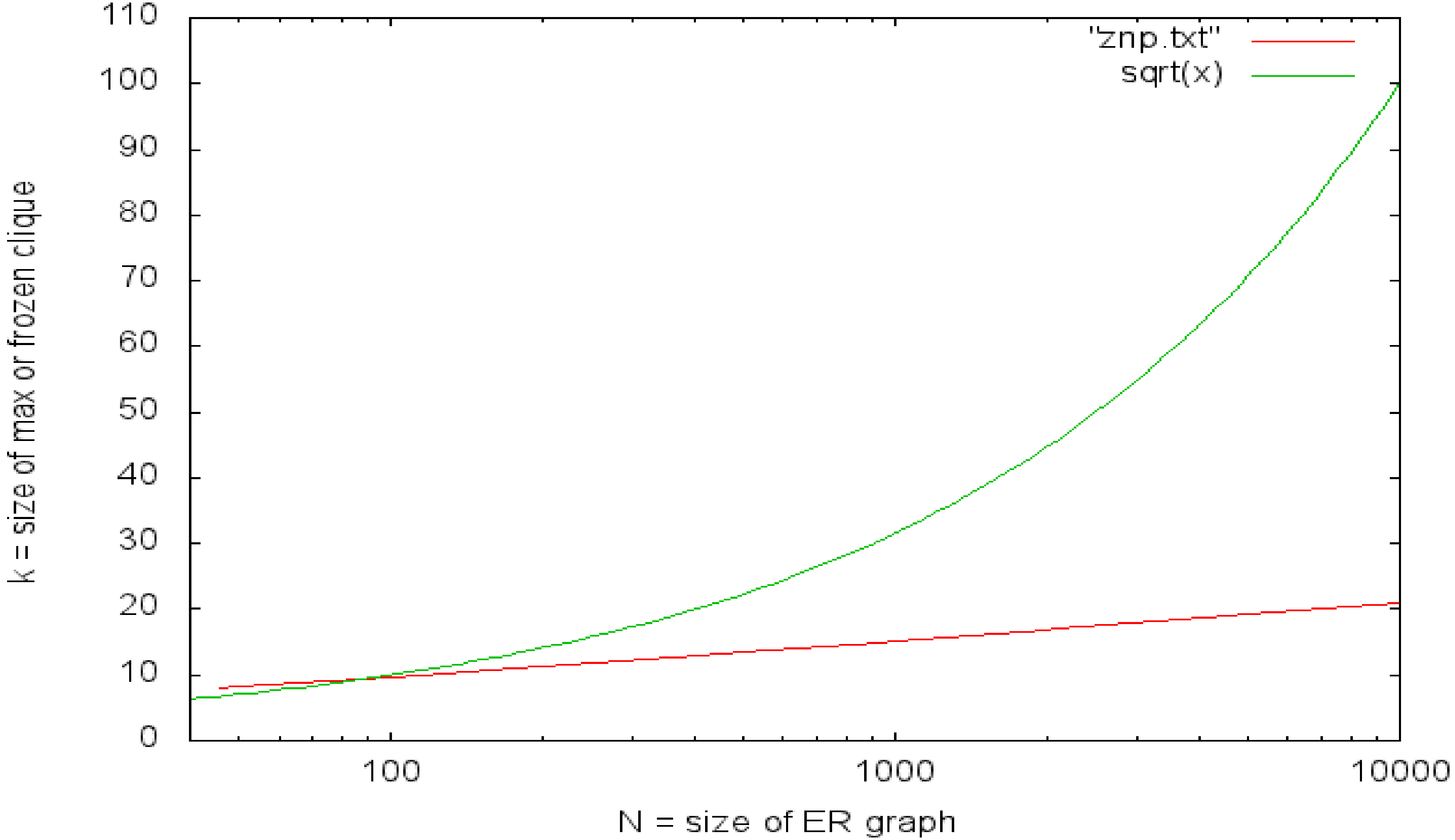
Other Formats: Paperback

I wish I had never loaned this to one of my students long ago. (but there is a downloadable version scanned and on the web)

The model of interest – maximal cliques

- In an E-R graph ensemble, G(N,1/2) how large is the largest clique?
- One of the earliest “sharp” phenomena in combinatorics.
  - Matula 1970-76 showed that it tends sharply to the integers closest to
    - R(N) = 1 + 2 log N = 2 log log N + 2 log (e/2) .
  - But this is just probabilities; how do you find it?
- Following old joke about mathematician and lamppost, much effort has shown that there are good ways to find planted cliques of order
  - c$\sqrt{NlogN}$ Kucera (1995)
  - or $c\sqrt{N}$ where C > 10 Alon, Krivelovich, Sudakov and many others
    - (note that any smaller fixed c can be managed, but things tend to get exponentially polynomial)
  - Can we adapt the tools used to search for even smaller naturally occurring cliques?
  - Deshpande and Montanari (2013) replace spectral approach with power method, acceleration tricks…

Phase diagram separating problems



Naturally occurring cliques are hard to find

- Lower bounds are weak
- Ramsey R(k,k) is proportional to log N, not 2 log N
  - R(k,k) $\leq \binom{2k-2}{k-1}$
- Greedy random search gives also log N, not 2 log N
- Simulated annealing has been “proven” to fail
  - Mark Jerrum, 1992
  - Naturally, this got my attention… but it is a more limited claim.

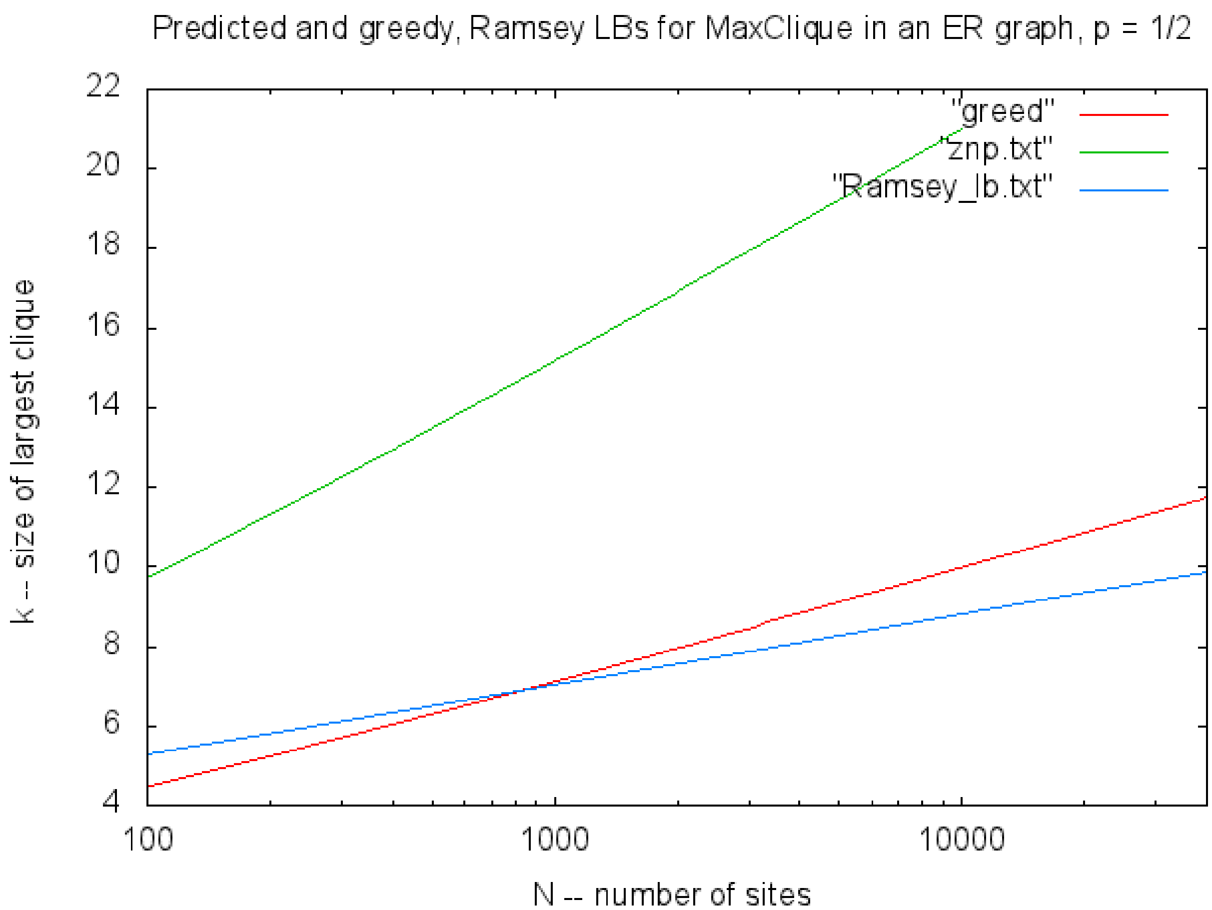
Large Cliques Elude the Metropolis Process

Mark Jerrum
Department of Computer Science, University of Edinburgh, The King's Buildings, Edinburgh EH9 3JZ, United Kingdom

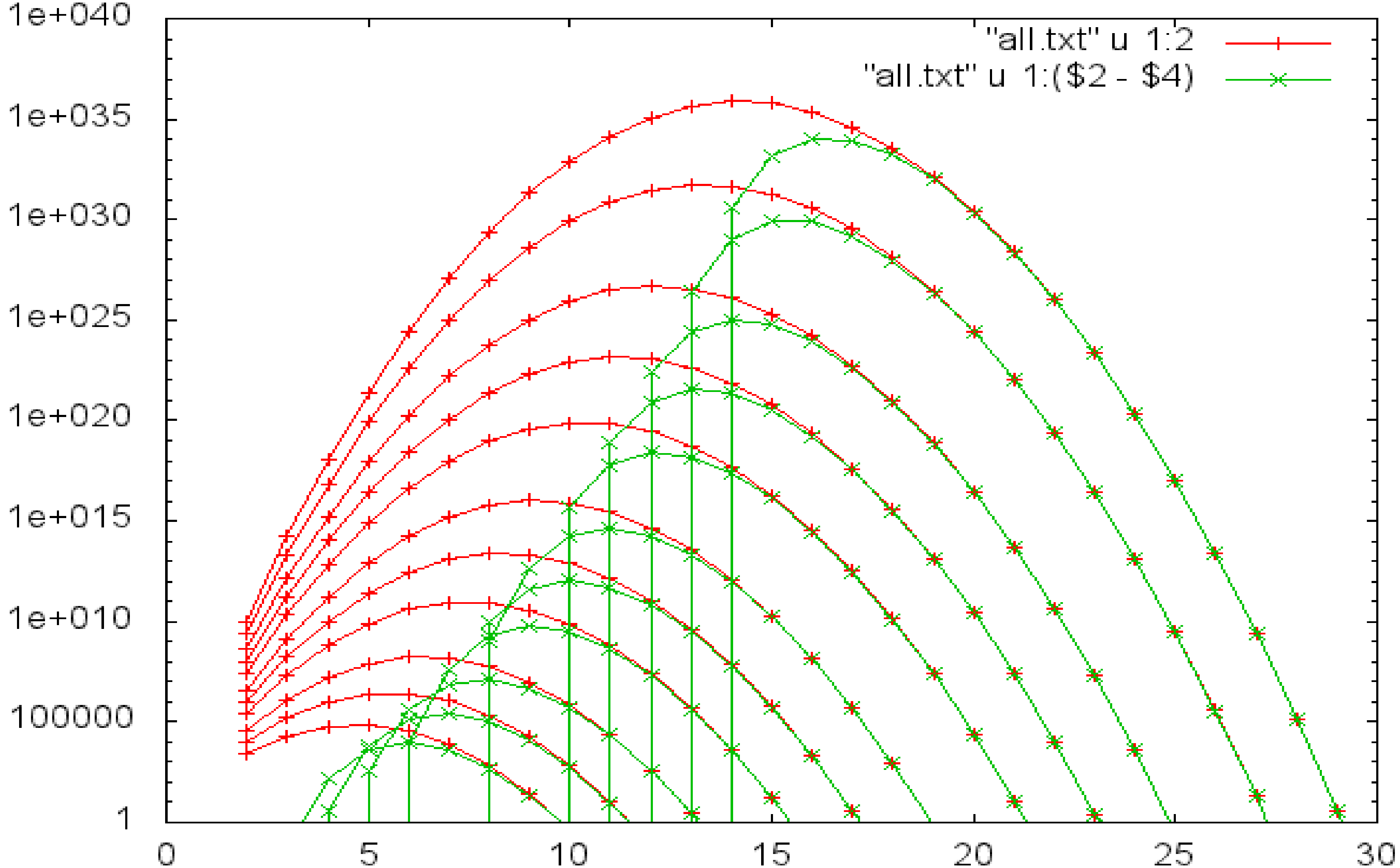
ABSTRACT

In a random graph on n vertices, the maximum clique is likely to be of size very close to $2 \lg n$. However, the clique produced by applying the naive "greedy" heuristic to a random graph is unlikely to have size much exceeding $\lg n$. The factor of two separating these estimates motivates the search for more effective heuristics. This article analyzes a heuristic search strategy, the *Metropolis process*, which is just one step above the greedy one in its level of sophistication. It is shown that the Metropolis process takes super-polynomial time to locate a clique that is only slightly bigger than that produced by the greedy heuristic.

The limits


Predicted and greedy, Ramsey LBs for MaxClique in an ER graph, p = 1/2
k -- size of largest clique
N -- number of sites
"greed"
"znp.txt"
"Ramsey_lb.txt"
4
6
8
10
12
14
16
18
20
22
100
1000
10000

Are cliques rare? Hardly!



Clique size k

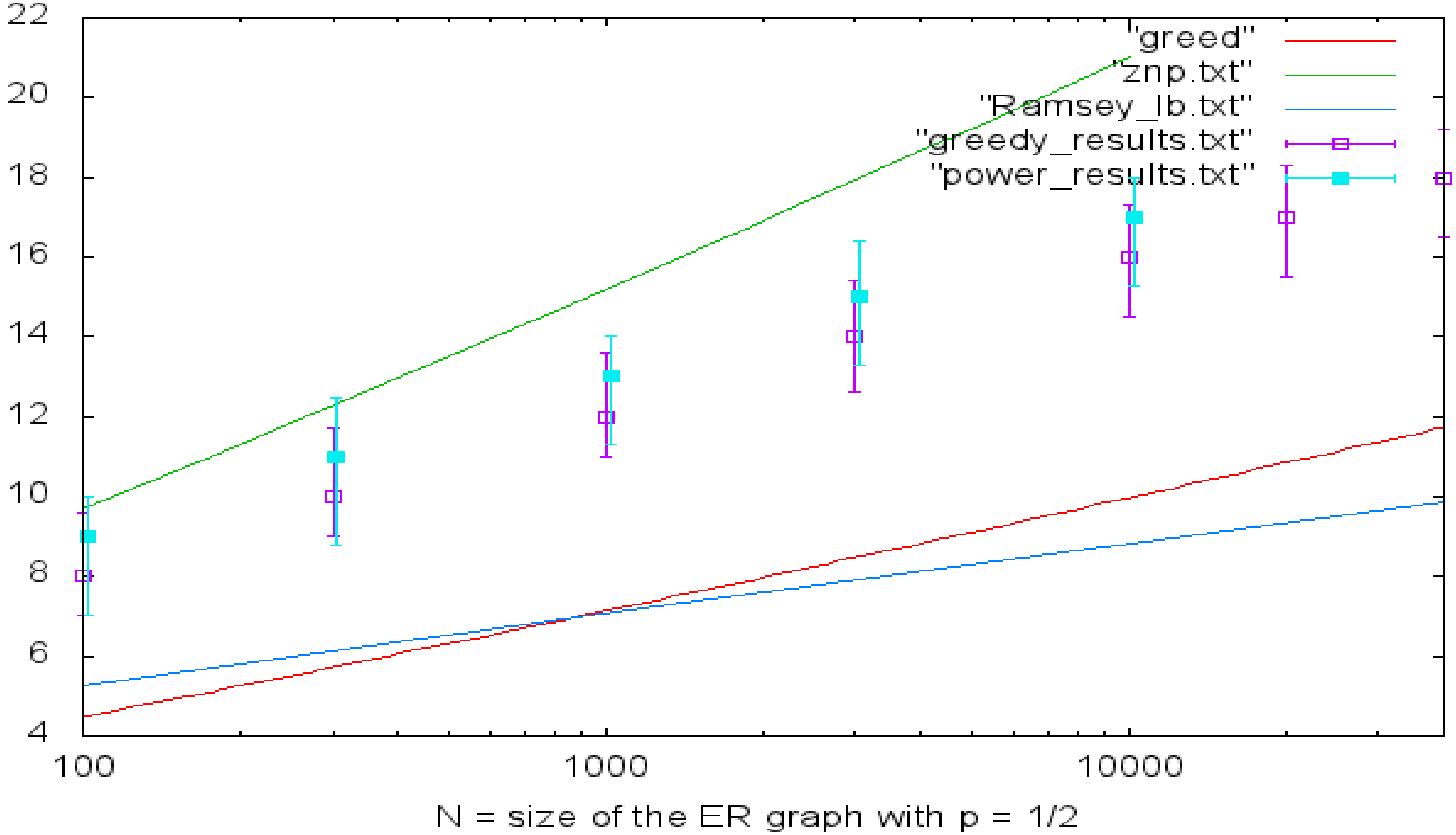
Material from the comments

- I'll put equations on the board for #cliques and for # cliques that cannot be extended further. Note many solutions are to be expected. And they seem to be distinct because k #(maximum cliques) = # (cliques of size k-1 that can be extended to size k). This implies the largest possible k-cliques are disjoint, making them possible to hit, since there are many…

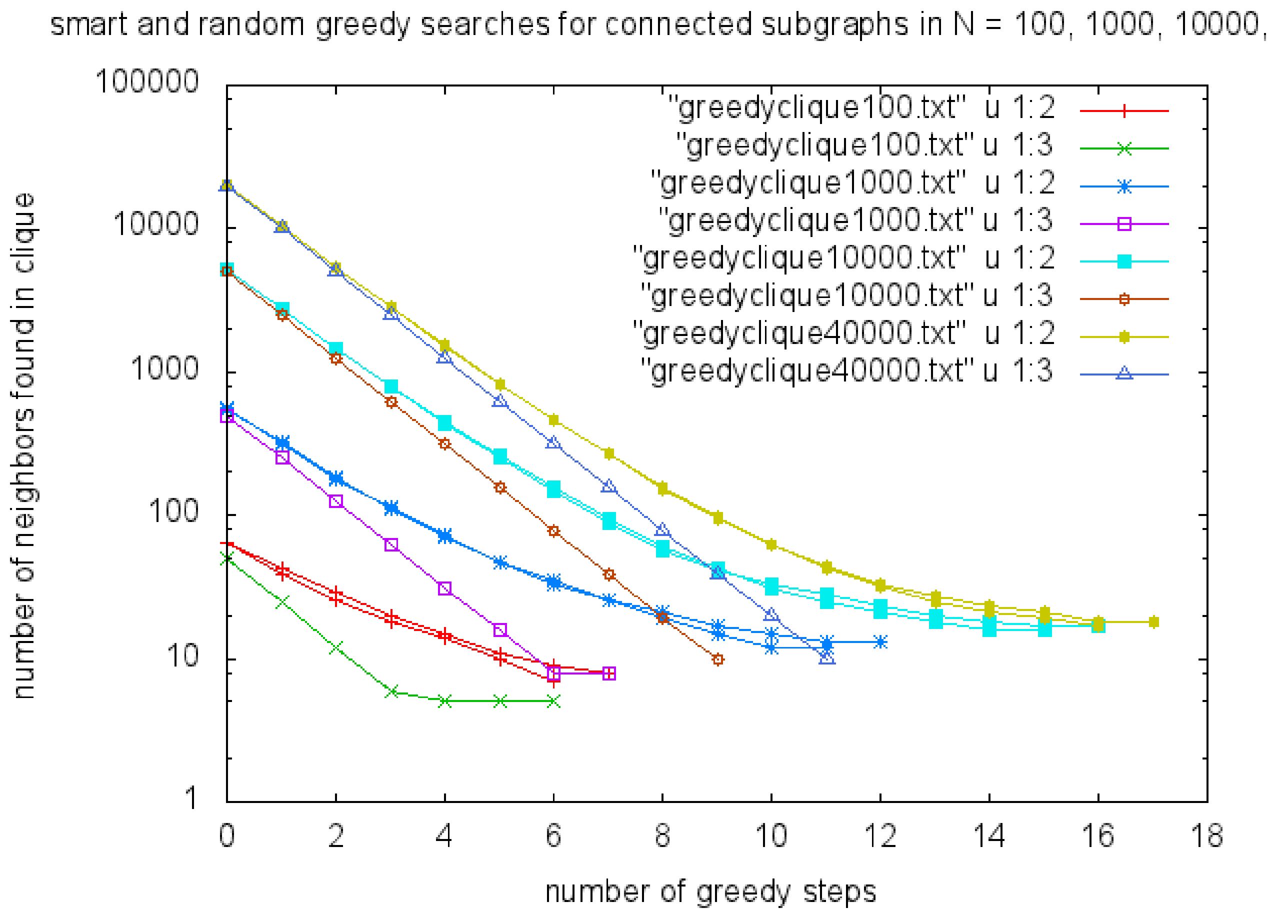
Improved search methods

- Smarter greedy – at each step select the site with max degree.
- Incorporate power method, at each stage in the restricted domain.
- De-anchor the search at intermediate or final stages and use power method on existing “friends” and their “family” (work in progress).

For natural max-clique, there are improvements, but problem is still not solved



Where were mistakes made?



A few lessons…

- Communities of interest can be local; modularity and dendrograms obscure this. Privacy and security studies also require high density or clique subgraphs.

- Power method and message-passing implement spectral imperatives in global scale networks – e.g. Page Rank. They offer soft tools for scaling search down.

- Iteration, or stochastic search always helps.