“ICTR Cloud Efforts”

developing “canonical” SIGINT analytics, finding hard targets and exploratory data analysis at scale

Data Mining Research – ICTR, GCHQ

Dr [redacted]

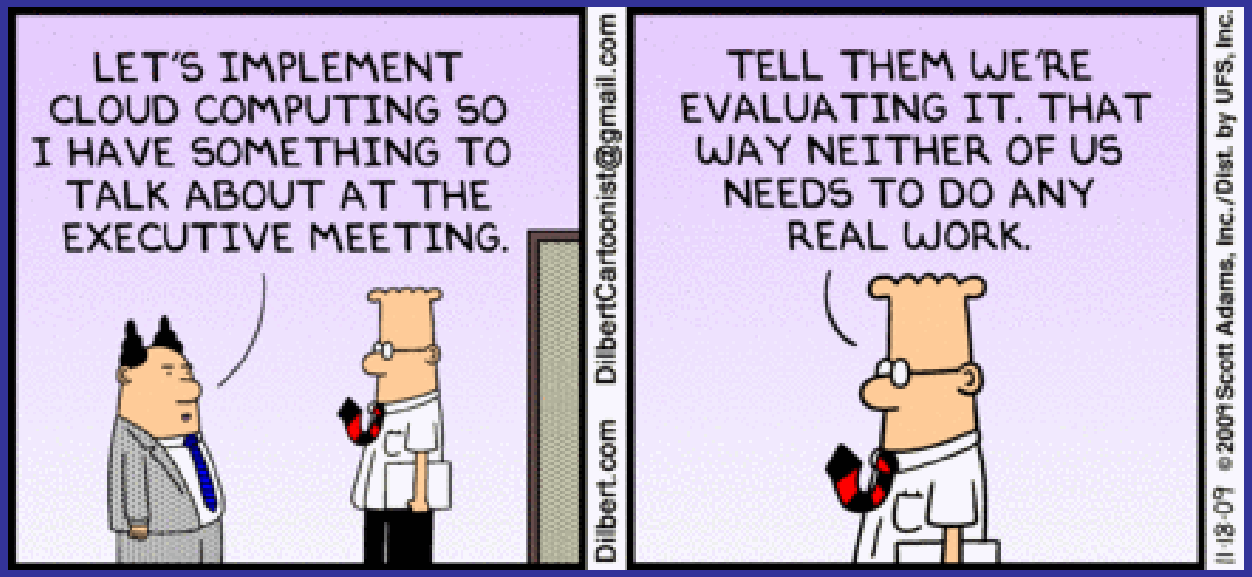
Building a SIGINT toolbox for BIG DATA

- Cloud analytics for SIGINT canonical operations
  - Aggregation - *building Geo-Time profiles for Internet Presence*
  - All pairs association – *alternate identifiers and Geo Associates*
  - Componentisation – *identify interesting small or large groups*

- Target discovery at population scale
  - *target discovery* – discover *unknown* targets
  - *known target communications behaviour* – modus operandi (MO)
  - *population scale* bulk *unselected* events - *all* events for country or world

- Exploratory Data Analysis of Internet / Cyber Events

GCHQ Cloud Analytic Development



www.dilbert.com/strips/comic/2009-11-18/

In last few years Data Mining Research at GCHQ have:

- developed new population scale analytics for multi-petabyte cluster
- evaluated cloud for data marting, bulk association, graph analytics
- delivered **operational benefit** – population scale target discovery

Geo-Summaries for all Internet presence

- Building Geo-Time profiles for *every* Internet identifier we see
- Discovering targets using Modus Operandi
- Summarisation of "Geo Pattern of Life" for every Internet identifier
  - Summarises how often each identifier seen in every country per week
  - Massively reduces data volumes (*trillions* of events to *billions* of profiles)

Email=[redacted]

Seen in: PK 17 times, UK 2 times

Week commencing	Seen
29/06/2009	UK,1
06/07/2009	UK,1
12/10/2009	PK,9
19/10/2009	PK,8

Perfect for MapReduce

IP-Geo for all Internet presence

Note scale of resulting profiles …

Geo-profiling over all presence events

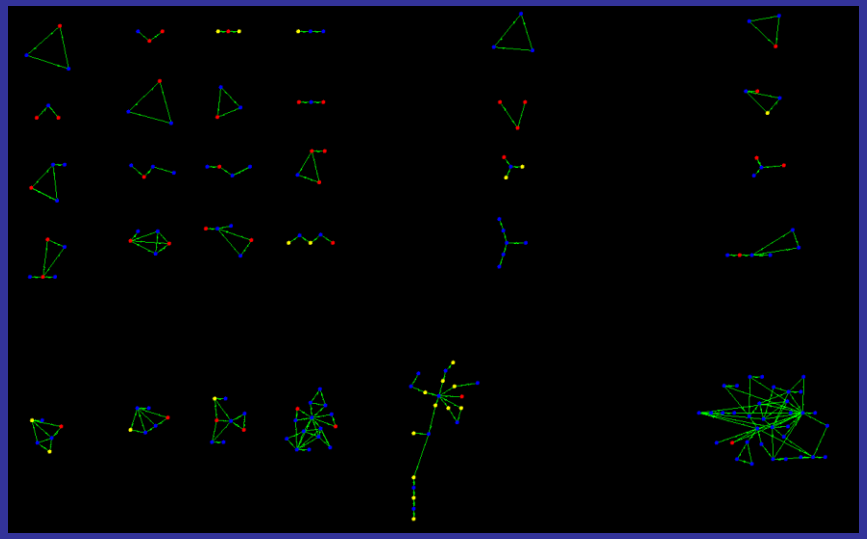
- **Perfect for MapReduce**
  - counting the number of occurrences in a large collection of records
  - ***"MapReduce: A Flexible Data Processing Tool"*** Dean and Ghemawat *Comms of the ACM* January 2010 **53**(1) pages 72-77

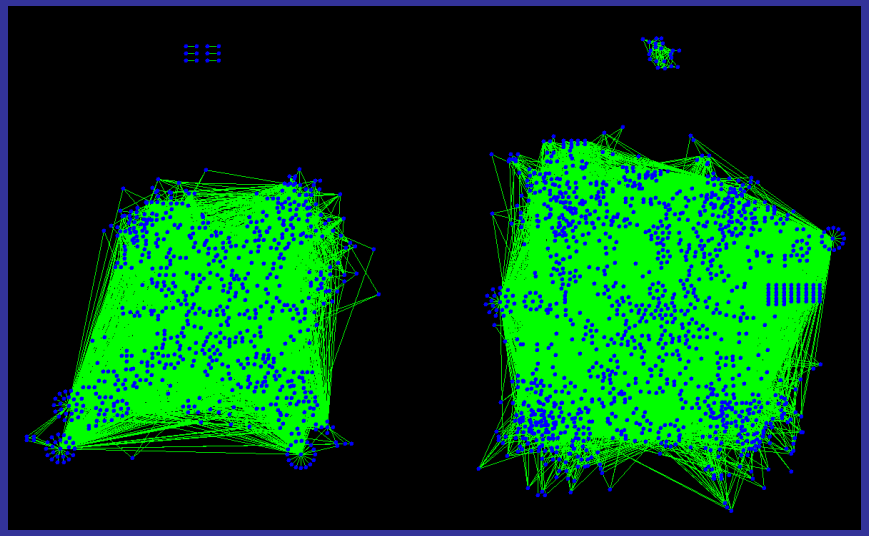
- The Geo-Time summaries for all target identifiers can be used to answer a number of questions:

  - *Where has this target identifier been?*
  - *Which target identifiers match the following country travel pattern?*
  - *Do anomalous Geo sightings indicate coordinated activity?*

- When combined with domain knowledge, can be extremely powerful if aggregated over all the data

EVERY ASSOC & BotGraph:
bulk pairwise associations and graph componentization



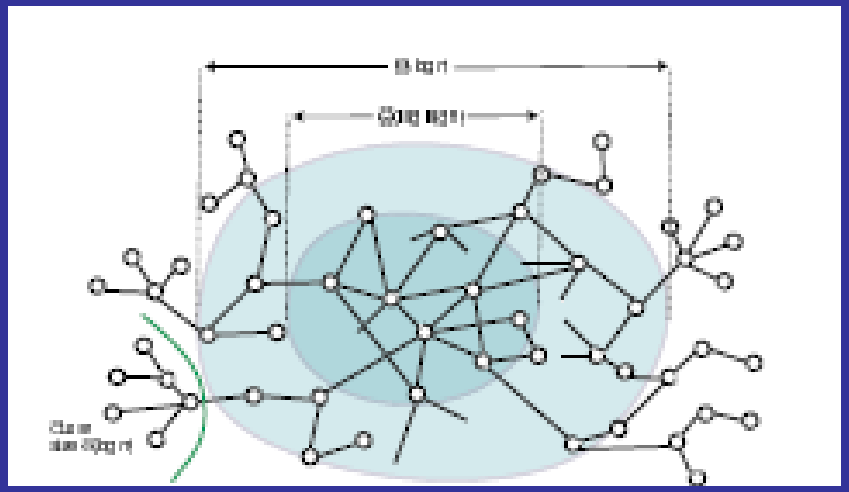


Large-scale community detection toolbox

- All pairwise correlation/association – build your graph
  - EVERY ASSOC for TDI alternate identifier scoring
  - BotGraph for webmail spam – Zhao et al *Botgraph* **[NSDI 09]**
  - PROBABILITY CLOUD for handset Geo-Association scoring
- Graph Componentisation
  - GCHQ MapReduce or Bagel implementation
  - Open source MapReduce implementations (CMU Pegasus)
- Analysis pattern to identify sub-sets for deeper analysis
  - Simple approach to make sense of huge datasets
  - Detect communities of potential interest from massive datasets
  - Rarely sufficient but essential first step in data volume reduction

Large networks are dominated by Giant Connected Component: *this can help you*

Leskovec, Lang, Dasgupta and Mahoney *Community Structure in large networks: Natural cluster sizes and the absence of large well-defined clusters* arXiv:0810.1355 (2008)



- Giant connected component dominates large networks
- Loosely connected periphery
- Relatively small number of disconnected small components

Target Discovery at Population Scale

- We are describing a *target discovery* technique based on *known target communications behaviour* applied to *population scale* bulk *unselected* events

- *target discovery* – discover *unknown* targets
- *known target communications behaviour* – modus operandi (MO)
- *population scale* – *all* the events we have for a country
- *unselected events* - not seeded on targets

Caveat Emptor

- Method has shown promise to discover phone groups of interest undiscoverable by traditional analysis.

- *"Find adversaries through their behaviour"*

- Initial identification of candidates *is* pure target discovery
  - not seeded on targets
  - search for behaviour in massive events

- ***BUT*** **it can only be used to effect if it is tied in with analyst knowledge of other patterns of behaviour, possibly geo-related.**

Critical Success Factors

- *Technical expertise in data mining (ICTR)*
- *Good understanding of target MO and ability to follow up new leads which are generated (Ops CT Analysts)*
- *Supporting IT infrastructure (SILVER LINING)*
- *Bulk access to relevant data sets (SILVER LINING)*
  - *ICTR lacks bulk access to CULT WEAVE – had snapshot in 2007*
  - *There were promising research lines: see SD conference 2007*

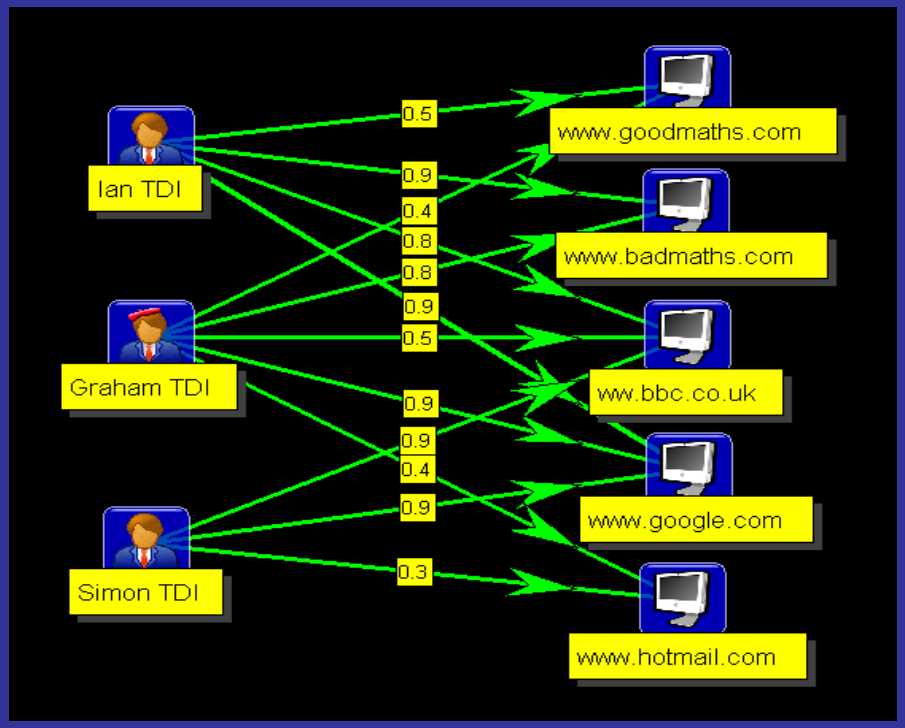
Operational Data Mining - *Key message*

- A combination of technical data mining experts, SIGINT developers, Operations analysts, appropriate data access and suitable IT is needed to make target discovery happen

- In our experience to date, it's not about tool development but the development of new (and fragile) data mining techniques by a critical mass of suitably skilled people!

- There are a set of cloud analytics that should form part of a toolbox but even then their successful application is likely to be as a result of collaboration with analysts

KARMA POLICE – *correlation between websites and internet IDs*



- Internet ID – IP – Web address: *correlation scored on statistics of IP*
  - KARMA POLICE QFD from ICTR
  - EVERY POLICE QFD on cloud
- Internet ID-website correlations form a weighted bi-partite graph
  - Links are weighted by KARMA POLICE correlation scores
  - Example graph showing correlations between Internet IDs and websites

AWKWARD TURTLE – *Cloud QFD*

- What is a recommender system?
  - Netflix – subscribers who like film X also like film Y
  - Amazon – customers who like book X also like book Y
  - GCHQ – Terrorists who like website X also like website Y

- MapReduce – vector of TDI scores for every website
  - Vector dot product – *"cosine similarity" measure*
  - Maximum degree TDI cut-off
  - Target activity is being used as similarity measure

- Website-website correlations – *found previously unknown file hosting*

Recommender Systems

- We have currently only used very simple techniques
- Body of active research
  - Netflix prize stimulated ☺
- Interested in seeing more statistical inference and large-scale modelling
  - Potential for long term research
- Behavioural targeting
  - Cf Google and Yahoo ad serving to subscriber profile

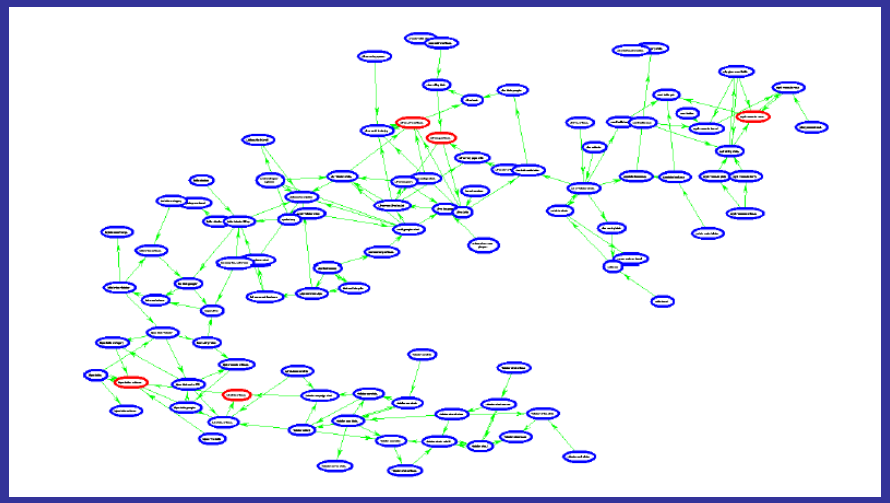
Query term graph

- Given a search term, which other search terms are related?

- Build Query term graph (MapReduce):
  - Nodes are queries
  - Directed edges between nodes if a machine searches for one term then the other within 5 minutes
  - Edge weighted according to frequency of search pattern

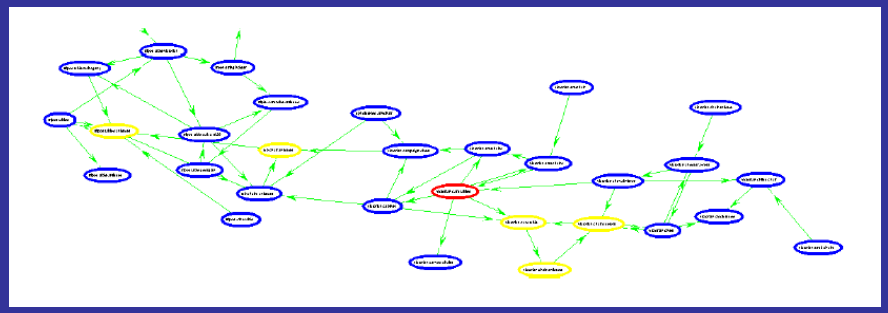
- Boldi, Bonchi, Castillo, Donato, Gionis and Vigna *The query-flow graph: Model and applications* **CIKM 08**
- Gionis *Efficient Tools for Mining Large Graphs* **MLG 10**

Ranking in Query Term graph - PageRank



- Small component from full query term graph
- All terms to do with different types of antiques
- Red nodes are top 5 PageRank scores

Personalised PageRank (PPR)



- Red node is seed node – *Victorian Card Tables*
- Yellow nodes are top 5 Personalised PageRank scores
- Nodes with high PR score also score highly with PPR

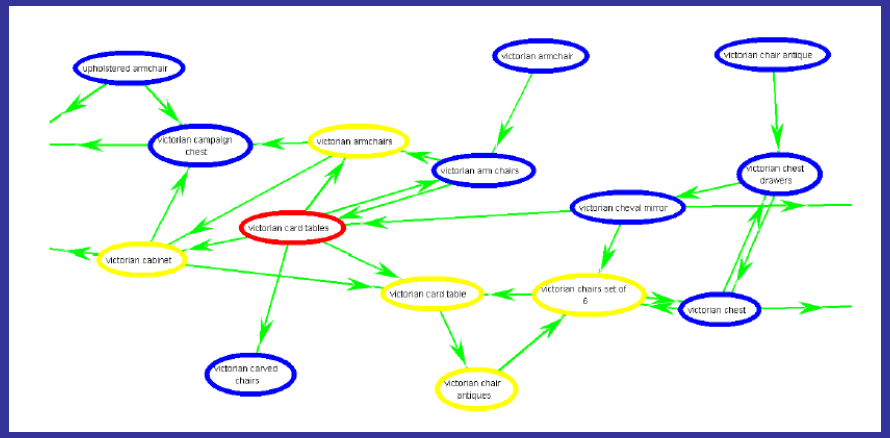
Normalised PageRank

- Want to find nodes with high Personalised PageRank score, $\mathbf{q}$, compared to its PageRank score, $\mathbf{p}$
- $\mathbf{p}$ and $\mathbf{q}$ are both (stationary) probability distributions on the same set so KL-divergence comes to mind

$$KL(\mathbf{q}||\mathbf{p}) = \sum_i q_i \log \frac{q_i}{p_i}$$

- We can rank the nodes based on their contribution to this sum, $q_i \log \frac{q_i}{p_i}$
  - This is the Normalised PageRank score

Normalised PageRank score



- Red node same seed
- Yellow nodes are top 5 Normalised PageRank scores
- Nodes with very high PageRank scores no longer dominate

Comments on Normalised PageRank

- Could go N-hops from seed node
  - Have to set pizza node degree limit
  - N-hop with pizza limit is standard contact chaining method
- Normalised PageRank deals with high degree nodes
  - High degree nodes tend to have high PageRank
  - Must score very highly on PPR to score well in Normalised PR
- Shown results within small component
- Evaluate Normalised PR for seed term in Giant Connected Component of Query Term Graph using Bagel

"GCHQ" seed query term

Rank	Query	NPR
1	Free People Check	0.791
2=	Jobs At Chanel	0.721
2=	Peter Wright (*Arabic*)	0.721
4	GCSE Bitesize Science	0.670
5	MI6	0.652
20	SKS	0.038
22	Foreign & Commonwealth Office	0.034
37	MI5	0.010
47=	MI6 James Bond	0.009
47=	MI^	0.009
47=	MI8	0.009
72	KGB	0.008
110	Wikileaks	0.003

Comments on Query Term Graph

- Query term graph is very noisy, as are all our Internet Events meta-data graphs
- Some promising results in finding similar queries but essential that results are interpreted by analysts
- Large amount of research to do
  - Clustering / Sessionising / … [lots of commercially motivated work]
  - Query chains – Banana -> Apple different intent to iPod -> Apple
  - Understanding the search behaviour of targets
- Normalised PageRank insights may be generally useful

Exploratory Data Analysis of Large-Scale Internet Events – *gap in understanding*

- Relevance to Cyber and SIGINT – *what is normal in the statistics of internet behaviour at large scale?*
  - Can we measure or model the salient features of large-scale internet communications meta-data?
  - Can we identify behaviours associated with target activity (be that human, machine or collective BotNet activity) that are detectable?

- GORDIAN KNOT (*Network Defence*) vs SIGINT feeds
  - Understand the potential of GORDIAN KNOT for Cyber EDA
  - What's the gap between GORDIAN KNOT and SIGINT data?

Internet/Cyber EDA – *FY 11/12*

- Fingerprint web browsing sessions
  - Can we ID a user based on their browsing habits?

- Is the Internet Regional?
  - Hypothesis: *"Internet is becoming more regionalised. Any machines communicating over long distances are of greater interest"*
  - Does the data support this?
  - Can we characterise the activity and significance of long distance communications?
  - Applications to Cyber, but also potentially to other Intelligence questions

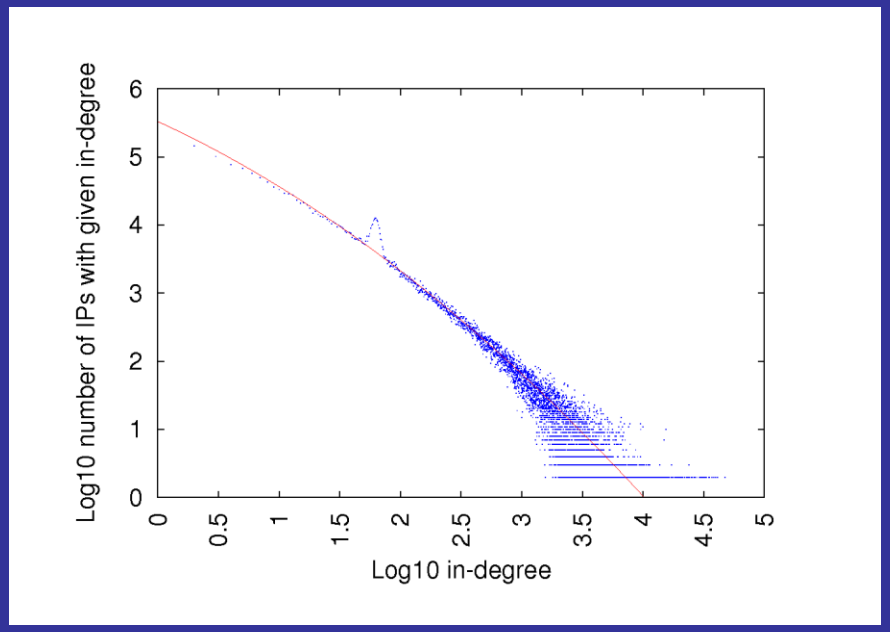
Internet/Cyber EDA – *FY 11/12*

- Attempt to identify malicious sites in the HTTP graph
  - *"BadRank* – given set of known *"bad"* web sites, can we identify associated sites that either point in same direction, or are reached from initial sites
  - *Identify loosely connected components* – bits that aren't closely tied in by association with Google et al.
  - *Subgraph detection* - if we have an approximate idea of how a user reaches a malicious web site, can we identify this pattern and similar others in the HTTP graph? [[redacted] SANDIA work]

Internet/Cyber EDA – *FY 11/12*

- FIVE ALIVE - *carry out EDA on the netflow dataset created by TR-FSP*
  - FIVE ALIVE is a bulk store of IP flow records, coupled with some very simple analytics that summarize and visualize IP activity
  - The main challenge here is to deal with the size of the dataset; current work in TR-FSP has revolved around looking at subsets of the data but it would be interesting to work on the dataset as a whole

BLOOD HOUND – *ICTR-NE*



- Detect electronic attack – aim to detect distributed and automated behaviour
- Idea from IDA/CCS SCAMP 2009
  - *'Using degree distributions to detect internet traffic anomalies*' Scheinerman
- Detect multiple IPs with same degree:
  - in-degree (*distributed hacking/port scanning*)
  - out-degree (*DDOS/bot tasking*)
- Graph: peak at in-degree ~ 10^1.8 = 63
  - Appears to be some sort of hacking activity
  - Dictionary attack: cycling through range of IPs on network, making 63 GET requests to each
  - Trying 63 combinations of URI, with the intent of getting a MySQL setup script (basic exhaust)

Summary

- Pattern-based data mining – unknown target discovery
  - Bulk unselected events – population scale – all events for country
  - Operational data mining – hard target discovery – real results
  - Target modus operandi – behavioural based discovery

- Selector-based data mining – unprecedented scale
  - Relationship scoring within multi-modal communications network

- Exploratory Data Analysis of Large-Scale Internet Events
  - *Gap in understanding of events at Internet Scale*
  - *How can BIG DATA analytics contribute to Cyber target discovery?*