Contact chaining

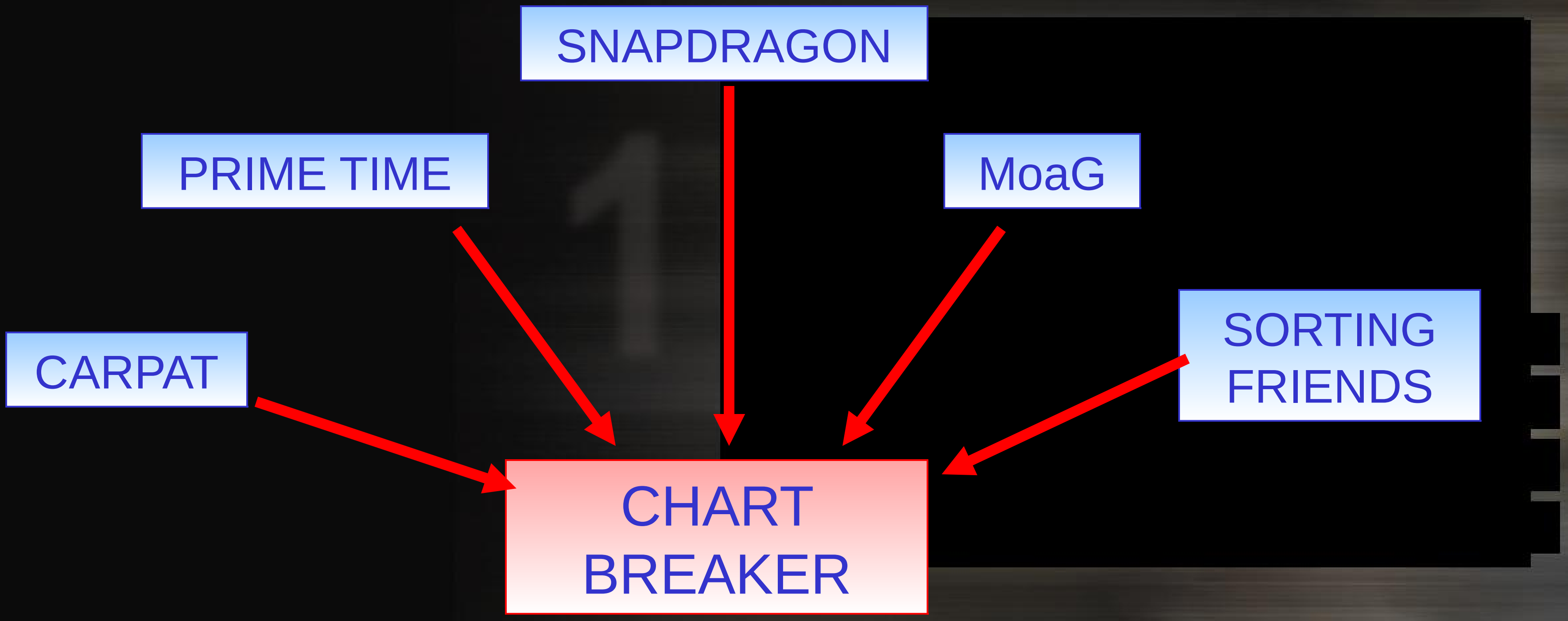
Task leader: [redacted]

Contact-chaining

- Start with known targets (comms identifiers)
- Expand data involving them to find new interesting comms identifiers
- B17 has worked on this for more than 8 years…

[redacted]

Contact-chaining:
Past prototypes



SNAPDRAGON

PRIME TIME

MoaG

CARPAT

SORTING FRIENDS

CHART BREAKER

[redacted]

Contact-chaining:
Drivers of current work

- Scoring, not just filtering
- Combining different algorithms
- C2C-specific techniques
  - Graph representation – loss of information
  - Far more rubbish in C2C
  - Growth of events volumes
- *Question-based*, interactive analysis

[redacted]

Combining scores:
SORTING FRIENDS

- Clustering telephone numbers that have a strong relationship – extrapolating social relationships from events
  - Direct communication
  - Co-location relationships
  - Timing and sequence
  - Call backs
  - Similar number
  - Equipment sharing
  - Calling profile similarity
  - Etc., etc.

  (currently 12 relationships in total)

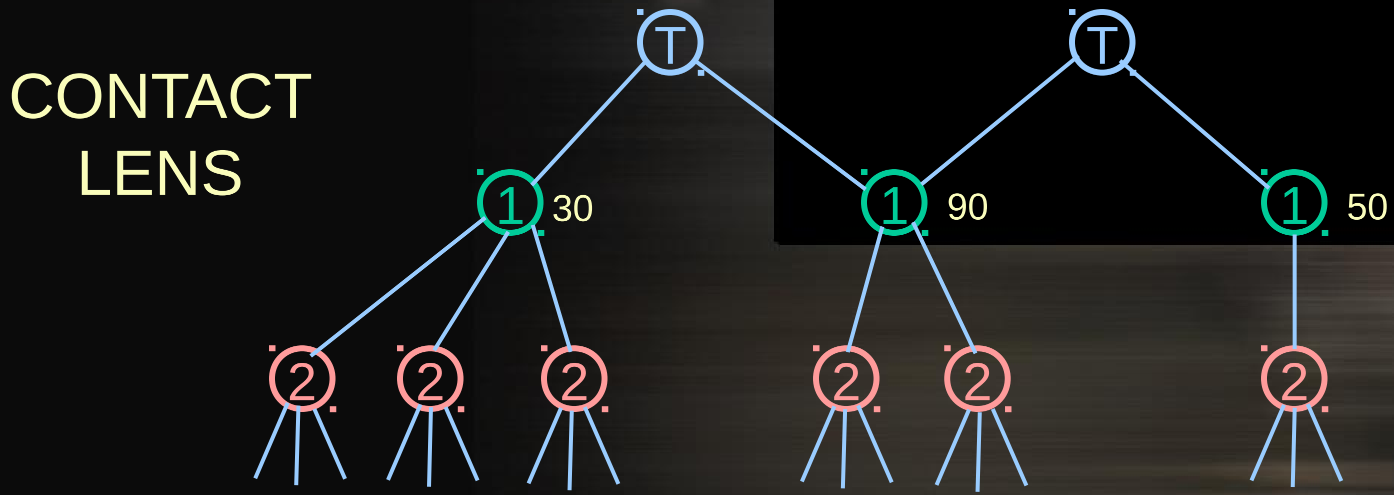
[redacted]

- Research prototype – SD can run… operational prototype will be through CHART BREAKER

Contact-chaining, problem 1: Data-mart generation

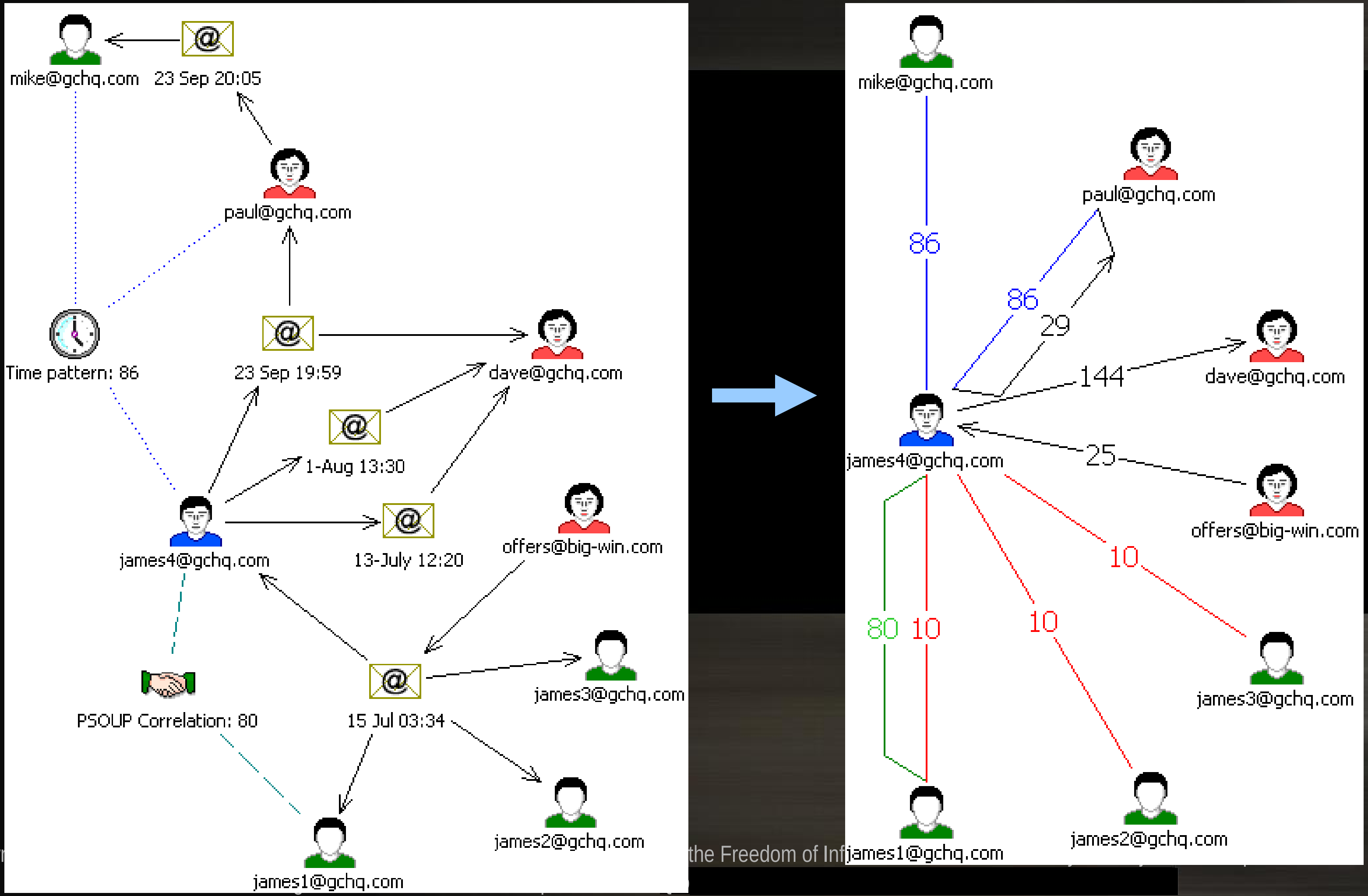
- Can no longer feasibly build a TNNN data-mart for a set of targets
- Streaming data-mart creation solves the problem
- But… quite complex, loses information and requires stable target set
- Instead use scoring to reduce work…

[redacted]



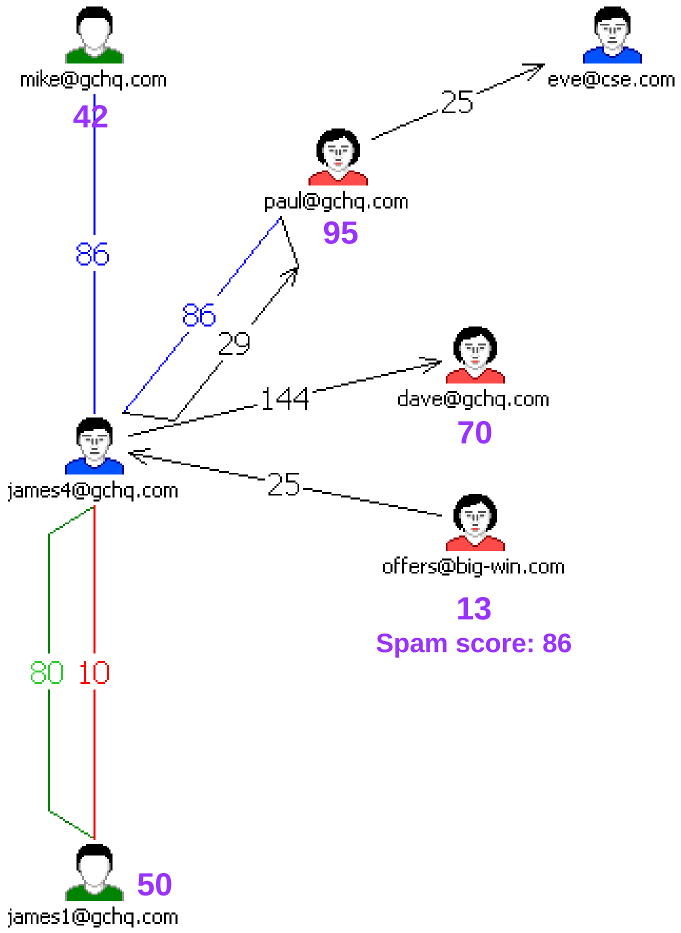
Contact-chaining, problem 2: Scoring edges

- Represent complex events as a graph, using edge scores:



Contact-chaining, problem 3: Scoring nodes

- Using total edge-scores as weights, score nodes to answer specific questions:



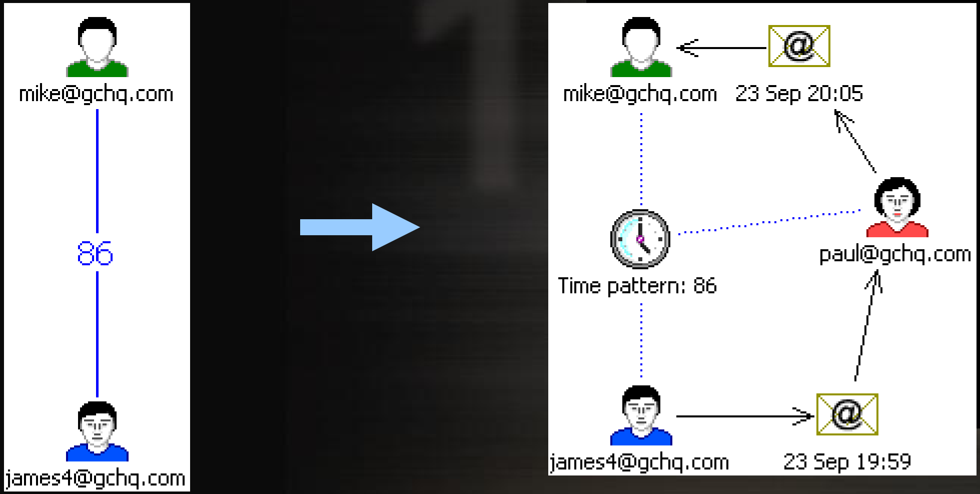
Most significant 5 non-targets, connected to your targets:

paul@gchq.com	95
dave@gchq.com	70
james1@gchq.com	50
mike@gchq.com	42
offers@big-win.com	13

Contact-chaining, problem 4: Interactive visualisation

- Multiple views – graph, lists of scores, textual information
- Descriptions of scores
  - X sent 2 emails to Y, both had no other recipients
- Ability to drill down on detail

[redacted]



- Interactive recalculation
  - E.g. non-target promoted to be a target

Contact-chaining:
CHART BREAKER

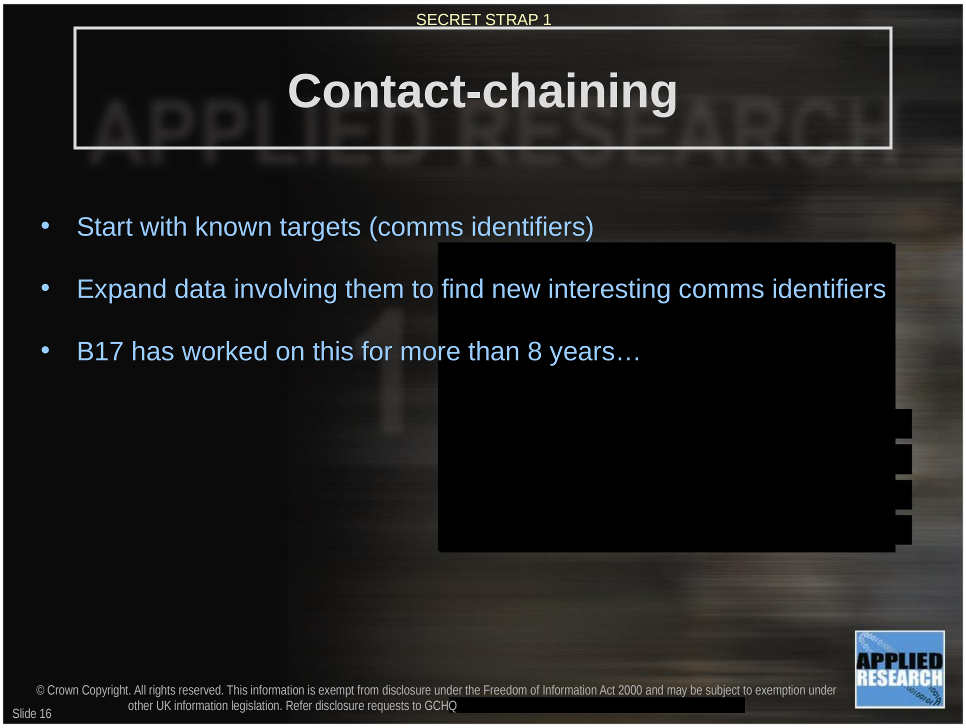
- Prototype software – based on Jung Java graph framework
- Flexible enough to quickly try different scoring techniques
- Uses CONTACT LENS to query Fleximart VLSP data
- Prototype visualisation will be through i2 / MAMBA

[redacted]

Contact chaining

[redacted]

Task leader: [redacted]



SECRET STRAP 1

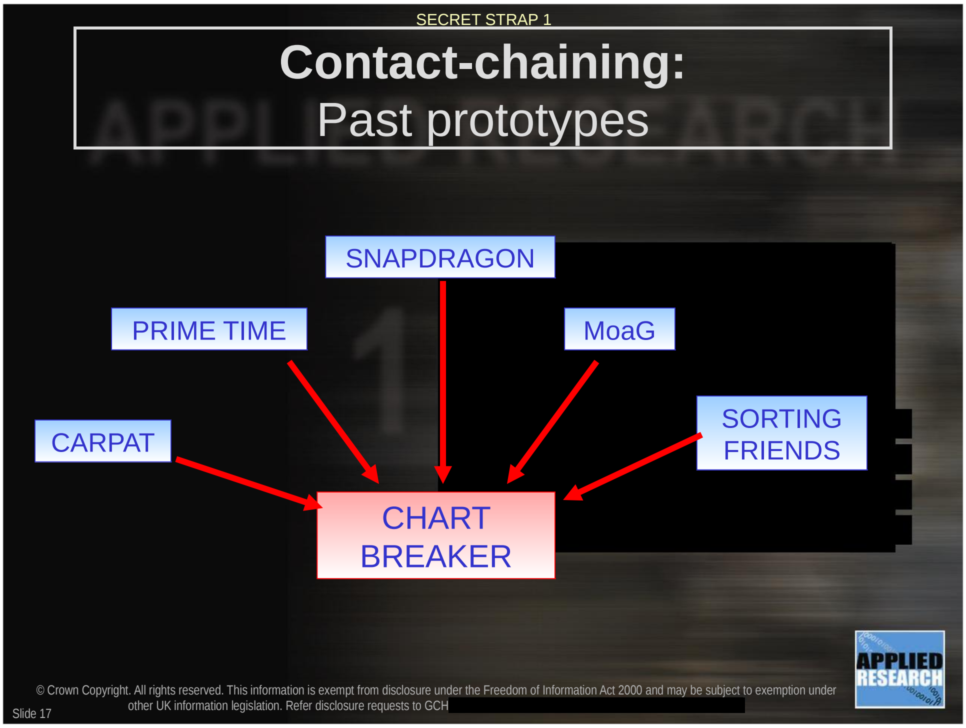
Contact-chaining

- Start with known targets (comms identifiers)
- Expand data involving them to find new interesting comms identifiers
- B17 has worked on this for more than 8 years…

[redacted]

Contact-chaining research is all about starting with a known set of targets (usually comms identifiers), and then looking at the data surrounding them to learn more or discover interesting new targets.

B17 have been doing research on this topic for at least 8 years!



SECRET STRAP 1

Contact-chaining:
Past prototypes

SNAPDRAGON

PRIME TIME

MoaG

CARPAT

SORTING FRIENDS

CHART BREAKER

[redacted]

Slide 17

There have been a whole load of past prototypes

CARPAT

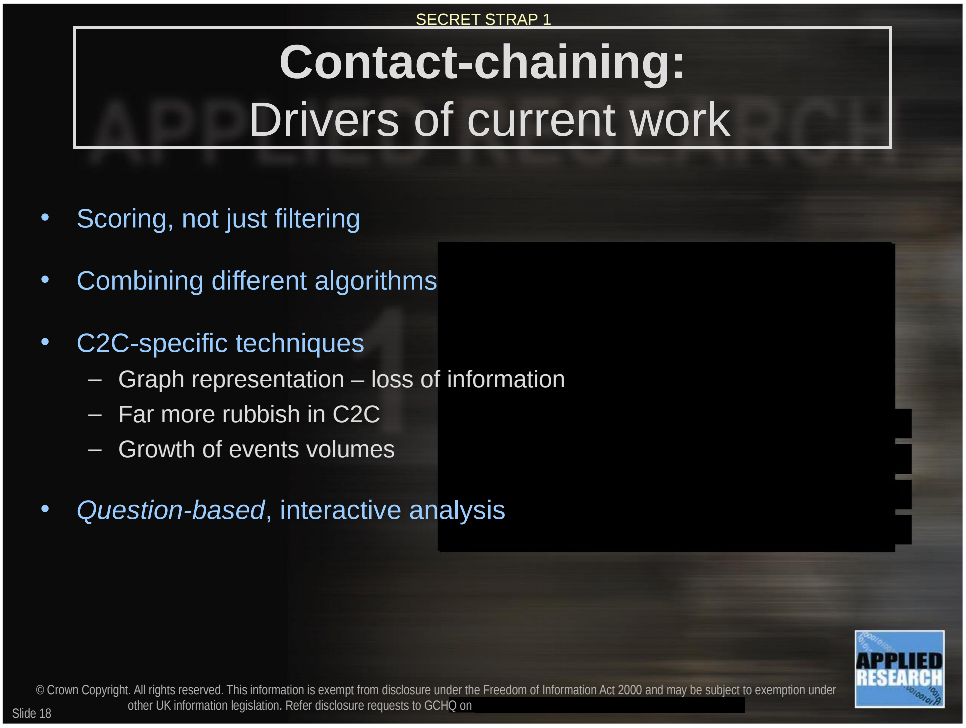
PRIME TIME – telephony analysis timing tool (which I will talk about later in the section of collaborative projects)

SNAPDRAGON – a C2C contact-chaining tool didn't work as intended

MOAG – our contact-chaining tool arising from the London bombings work, which is still being used regularly

SORTING FRIENDS – another telephony contact-chaining tool developed by an integree at NSA.

The contact-chaining task is taking all the lessons from these and will eventually create a new prototype, CHARTBREAKER.



SECRET STRAP 1

Contact-chaining:
Drivers of current work

- Scoring, not just filtering
- Combining different algorithms
- C2C-specific techniques
  - Graph representation – loss of information
  - Far more rubbish in C2C
  - Growth of events volumes
- *Question-based*, interactive analysis

[redacted]

APPLIED RESEARCH

Slide 18

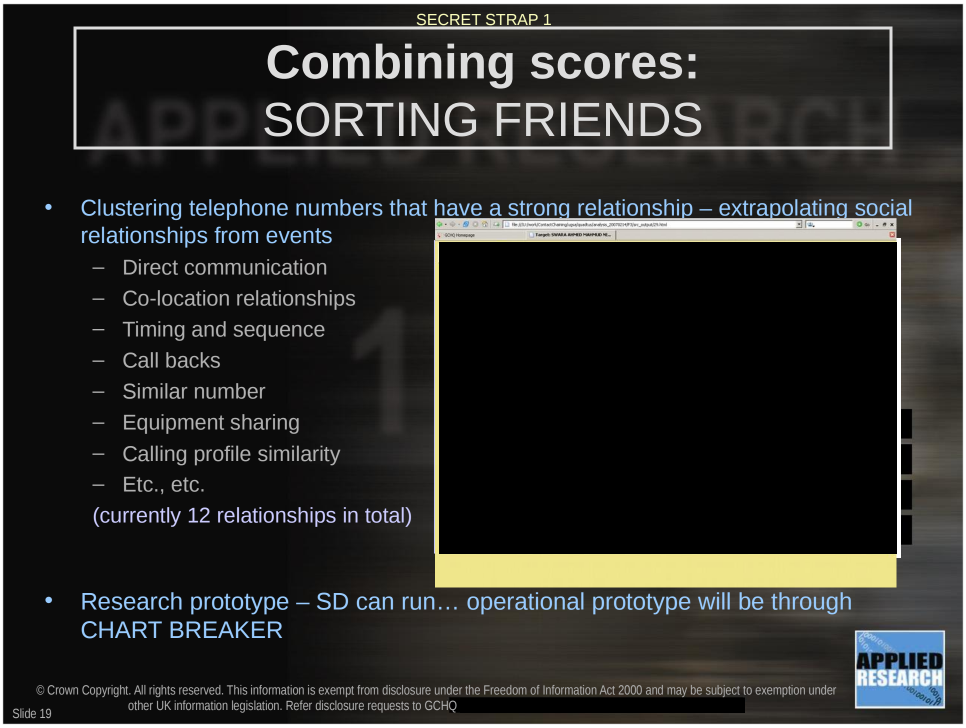
The main themes of our contact-chaining work are:

-Scoring, not just filtering – CARPAT is a filtering tool, it takes bulk data and filters to get the n/w surrounding targets. However, these networks can still be pretty big, and will get bigger as data volumes continue to grow. We think it is going to be important to be able to score nodes in these networks, so that users can home in on the most important bits

-Another important factor is combining together different algorithms and data sources without confusing the user.

-We also think it is important to develop CC algorithms specifically for CC as our recent experience is that taking an algorithm designed for telephony analysis and just porting to C2C data does not work very well. The reasons for this are:…

-Finally, we think it is important to make contact-chaining tools question-based and interactive. For example, one of the problems with many tools is that they present you will a whole load of configurable options and parameters, which can overwhelm the user. We think we should move to a situation where the user can ask a specific analysis questions of the tool – for example “who are the most strongly connected email addresses to this set of targets” or “what are the links between these two groups of targets?”. The tool presents the best answers it can, allowing the user to interact with the results, drilling-down on detail, in a similar way to IMMINGLE.



SECRET STRAP 1

Combining scores: SORTING FRIENDS

- Clustering telephone numbers that have a strong relationship – extrapolating social relationships from events
  - Direct communication
  - Co-location relationships
  - Timing and sequence
  - Call backs
  - Similar number
  - Equipment sharing
  - Calling profile similarity
  - Etc., etc.

  (currently 12 relationships in total)

file:///U:/work/ContactChaining/ugsa/quadtus/analysis_20070214/F3/src_output/29.html
Go
GCHQ Homepage
Target: SWARA AHMED MAHMUD NE...

[redacted]

- Research prototype – SD can run… operational prototype will be through CHART BREAKER

I mentioned combining of scores.

Good example is a telephony prototype, SF, developed between GCHQ and R62 at NSA (by [redacted], GCHQ integree there)

SORTING FRIENDS takes a large number of tel call records (e.g we've been running it on the A24 datamart) and reduces it down to all sorts of different relationships between phone numbers. The aim is to try and use lots of different measures of strength of relationship to get closer to the real social relationship.

Describe patterns

Then there is a clustering and visualisation step.

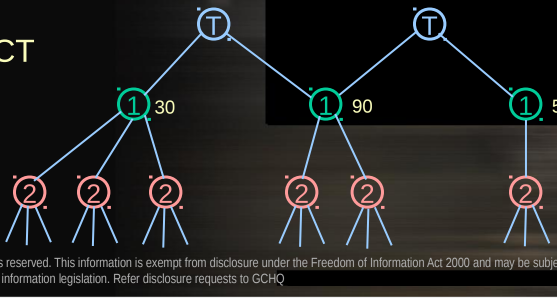
Currently SF is very much a research prototype, and Lorraine in SD can run it for large projects in Ops. Much work needed on it, we don't think it will be a pulled-through on its own, but rather will be part of CHART BREAKER.

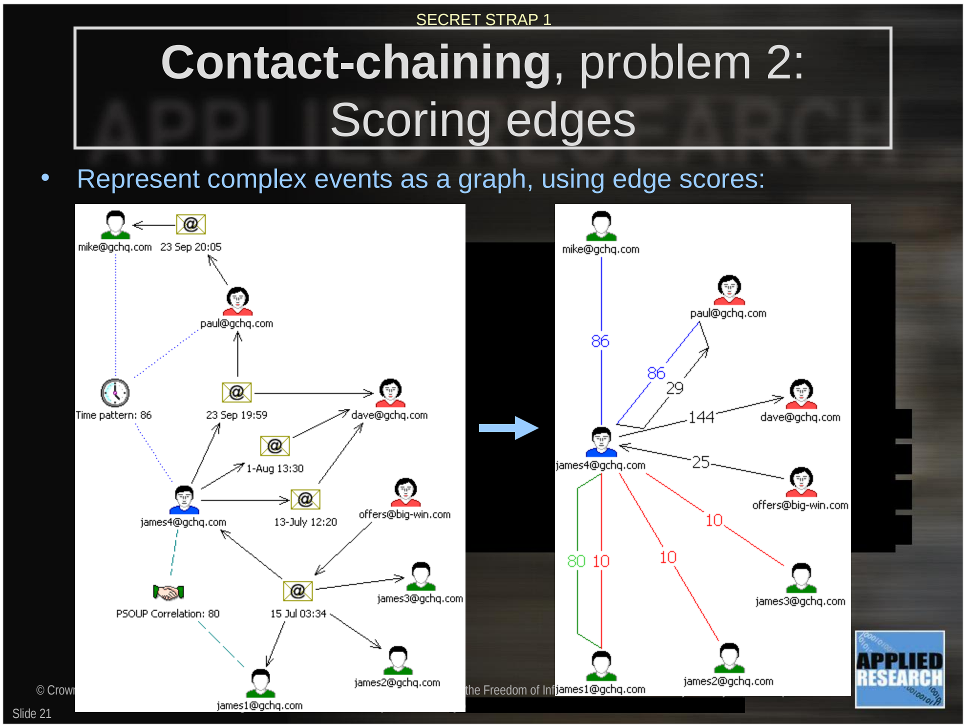
Contact-chaining, problem 1: Data-mart generation

- Can no longer feasibly build a TNNN data-mart for a set of targets
- Streaming data-mart creation solves the problem
- But… quite complex, loses information and requires stable target set
- Instead use scoring to reduce work…

[redacted]

CONTACT LENS





SECRET STRAP 1

Contact-chaining, problem 2: Scoring edges

- Represent complex events as a graph, using edge scores:

© Crow

the Freedom of Inf

Slide 21

Mentioned problem of not being able to represent full complexity of C2C events as a graph… Here's an example of a full network… emails and other things such as PSOUP score and timing pattern.

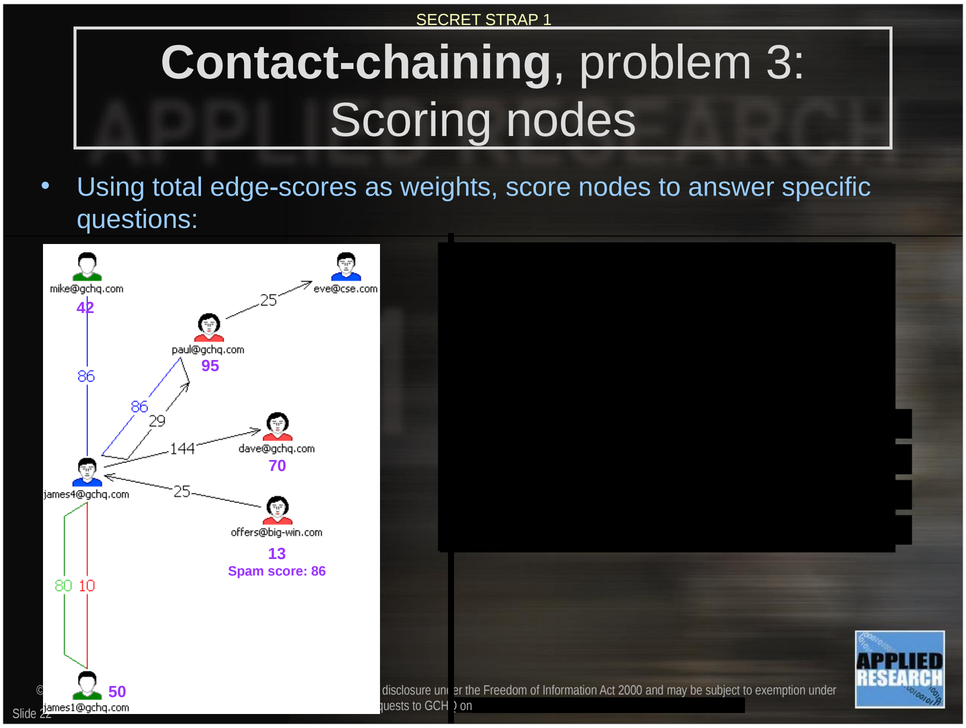
We think the key to this is to collapse down to edges, but with weights on those edges that measure how strong a relationship there is.

So the full network reduces to this graph – blue timing pattern, black sender-to-receiver, red relationship between 2 recipients, and green PSOUP.

Trick is to get the edge weights right – at the moment we're working on algorithms that can do this for all the relationships in email (sender-to-receiver, receiver-to-receiver, event sender-to-sender).

Also looking at timing analysis for C2C.

Where you have multiple types of relationship between people, you can also calculate a total score.



SECRET STRAP 1

Contact-chaining, problem 3: Scoring nodes

- Using total edge-scores as weights, score nodes to answer specific questions:

[redacted]

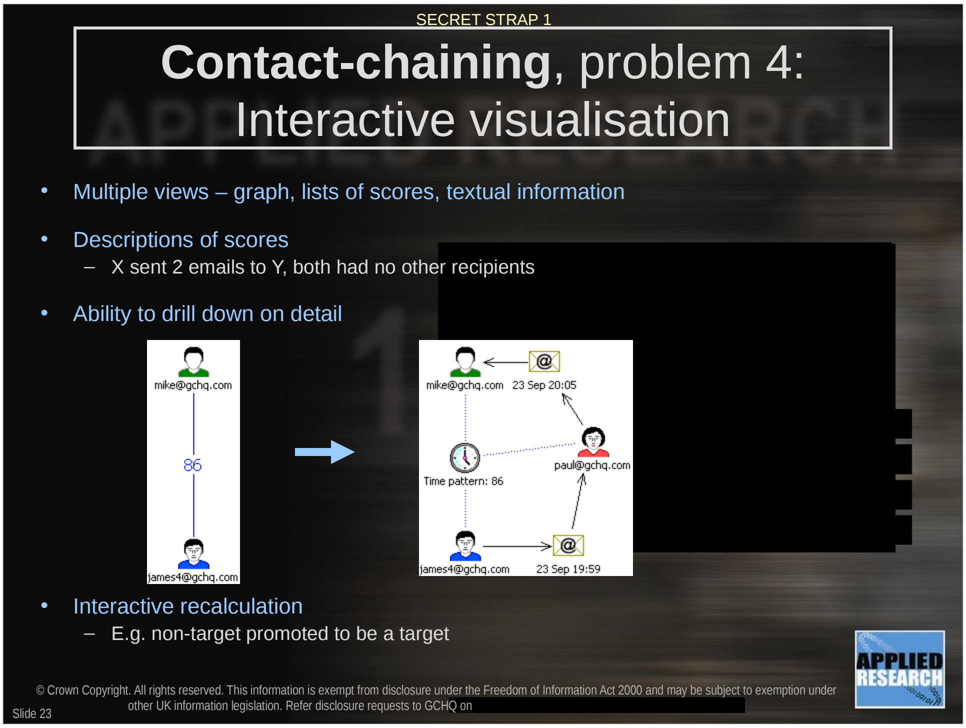
© disclosure un er the Freedom of Information Act 2000 and may be subject to exemption under quests to GCH) on [redacted]

Slide 22

After you have edge scores, you can use these to score nodes, to answer specific questions.

For example, in CHART BREAKER we have developed a new scoring technique that scores non-targets according to how strong the relationship is to a small group of targets… as shown here. This allows you to rank non-targets by suspiciousness. The algorithm uses a technique that gives targets a weight and sort of iteratively distributes it down the links to connected nodes over a number of iterations, (not just to first circle nodes). We want to develop similar algorithms for other questions. Node scoring algorithms may be independent of whether telephony or C2C due to the edge scoring stage.

This is where you include pre-calculated scores that concern nodes themselves – for example, offers@ has a spam score, which lowers score in answer to the question. The scores allow you to rank nodes according to how well they answer the question, and only show the top 5 scores.



SECRET STRAP 1

Contact-chaining, problem 4: Interactive visualisation

- Multiple views – graph, lists of scores, textual information
- Descriptions of scores
  - X sent 2 emails to Y, both had no other recipients
- Ability to drill down on detail

[redacted]

mike@gchq.com
86
james4@gchq.com
mike@gchq.com
23 Sep 20:05
paul@gchq.com
Time pattern: 86
james4@gchq.com
23 Sep 19:59

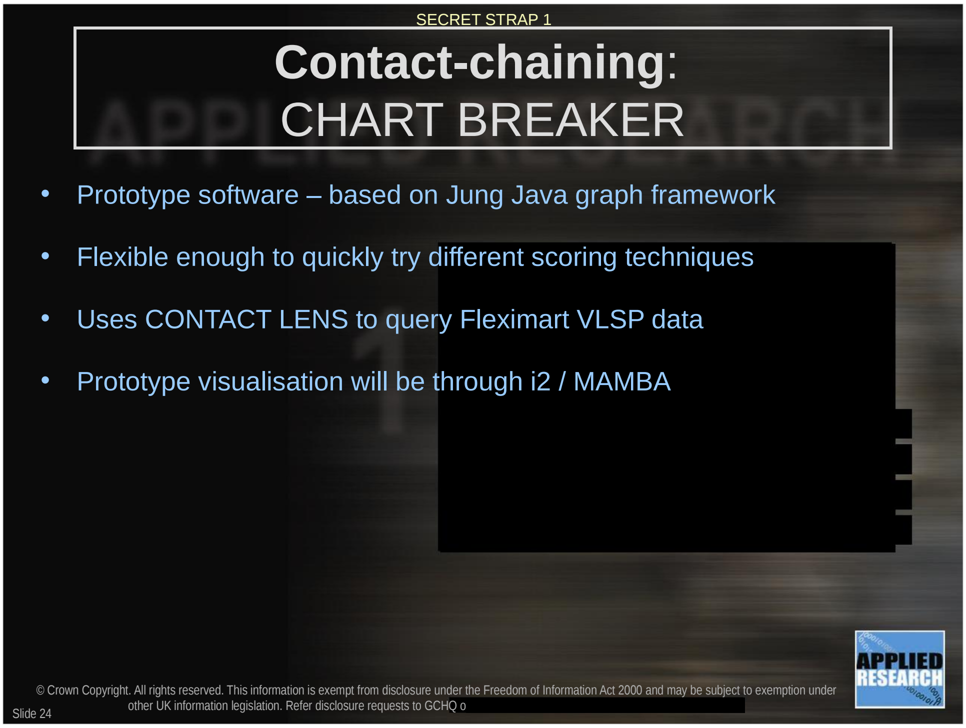
- Interactive recalculation
  - E.g. non-target promoted to be a target

We think the visualisation will be key to using contact-chaining algorithm.

We think it is important for the prototype to be able to show multiple views of the data….

Also, descriptions of "why"

For this reason, we've started some work on using MAMBA to visualise the results of contact chaining algorithms



SECRET STRAP 1

Contact-chaining: CHART BREAKER

- Prototype software – based on Jung Java graph framework
- Flexible enough to quickly try different scoring techniques
- Uses CONTACT LENS to query Fleximart VLSP data
- Prototype visualisation will be through i2 / MAMBA

[redacted]

Mention software will probably need to be rewritten to be more efficient once we know what we the scores are.