Data mining then, now, and next

David J. Hand

Imperial College London

and

Winton Capital Management

July 2012

‘the discovery of interesting, unexpected, or valuable structures in large data sets’

What is ‘large’?

Any data set which cannot be analysed by hand?

But ‘large’ keeps growing

Presenting new technical challenges

- of computational infrastructure
- of inference

Large keeps growing

'... a million, being a number so enormous as to be difficult to conceive.'

Francis Galton, p11 of *Hereditary Genius*

But nowadays ***millions***, ***billions***, and ***trillions*** are commonplace

10^6 10^9 10^{12}

Megabyte	2^{20} bytes ($\approx 10^{6}$ bytes)
Gigabyte	2^{30} bytes ($\approx 10^{9}$ bytes)
Terabyte	2^{40} bytes ($\approx 10^{12}$ bytes)
Petabyte	2^{50} bytes ($\approx 10^{15}$ bytes)
Exabyte	2^{60} bytes ($\approx 10^{18}$ bytes)

Drivers

Business

April 2011: 24 million plastic card transactions per day in UK

The natural sciences

100 petabytes of data from LHC has been analysed

The life sciences

“The study of genomics increasingly is becoming a field that is dominated by the growth in the size of data and the responses by the broader scientific community to effectively use and manage the resulting derived information ... 40 Gb per day on a singe sequencer, and there are now 10 to 20 major sequencing labs wordwide that have each deployed more than 10 sequencers.”

Science, 11 Feb 2011, 331, p728

The social sciences

2 billion people use internet
14% increase in 2010
(so all my numbers are out of date underestimates)

294 billion emails sent per day
107 trillion emails sent in 2010

100 million new Twitter accounts in 2010
25 billion tweets sent in 2010

6 billion mobile phone calls per day in the US

London has 6 million CCTV cameras

Capability of computer hardware increases by ***Moore's Law***:

'the complexity for minimum component costs has increased at a rate of roughly a factor of two per year'

$\Rightarrow$ data processing power increasing rapidly

But data acquisition technologies exceed this

- electronic/automatic data acquisition

More importantly (***Hand's Principle !***):

data processing is not data understanding

Two *perspectives* on data mining

Structures vs algorithms

Detecting structures in data
(statistics, machine learning)

Applying algorithms to data
(computer science)

Two *aspects* to data mining

Model building vs pattern discovery

Model building

- characterising large scale features of data sets
    - predictive models for large subpopulations
    - segmentation of entire data sets

Pattern discovery

- locating small departures, anomalies, peculiarities

Two *challenges* of data mining

Infrastructure vs inference

How do you search, sort, match, filter, etc when
- there are billions of data points
- the data keep on coming
- you need the answer *now* ?

How do you tell if a ‘discovery’ is due to
- chance
- poor data quality
- already known
- or plain uninteresting?

Data issues in data mining

Data set size

Data are often non numerical

text, images, waveforms, trajectories, graphs, ...

Data complexity

distributed, dynamic (individual and/or population),

Data quality

selection bias, missing values, ...

Data structures

complex

Examples of modelling tools

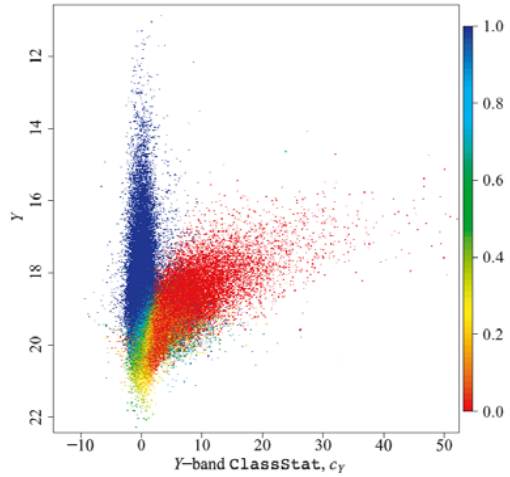
Clustering

An important distinction:

1) Segmentation: *carving nature at the joints*

Finding natural groupings

e.g. Classic early work (early days of computers, before the term ‘data mining’ became widespread) in determining types of depression: bipolar, unipolar



2) Dissection: *finding a convenient grouping*

e.g. 1: in an advertisement in the *Sunday Times* of 18th April 1999, James Meade Limited, a shirt manufacturer, gave a choice of sizes, as follows, where a ✓ denotes sizes available, and ✓ denotes standard size

	Sleeve length (inches)						
	31	32	33	34	35	36	37
Collar size							
14½	✓	✓	✓	✓	✓	✓	✓
15	✓	✓	✓	✓	✓	✓	✓
15½	✓	✓	✓	✓	✓	✓	✓
16	✓	✓	✓	✓	✓	✓	✓
16½	✓	✓	✓	✓	✓	✓	✓
17	✓	✓	✓	✓	✓	✓	✓
17½	✓	✓	✓	✓	✓	✓	✓
18	✓	✓	✓	✓	✓	✓	✓

e.g. 2: FRuitS (***F***inancial ***R***esearch ***S***urvey)

60,000 respondents to an NOP survey are classified into eight categories according to lifestage, financial strength, and product portfolio

	Financial Strength								
Savings	None	None	None	Moder	Moder	Moder	High	High	High
Income	Low	Ave	High	Low	Ave	High	Low	Ave	High
Lifestage									
1. Young-single	Lemon	Lemon	Grape	Orange	Orange	Orange	Orange	Orange	Orange
2. Single, 25-34	Lemon	Grape	Grape	Orange	Orange	Orange	Orange	Orange	Orange
3. Youngcouple	Lemon	Grape	Grape	Orange	Apple	Cherry	Orange	Pear	Plum
4. Young family	Lemon	Grape	Grape	Orange	Apple	Cherry	Pear	Pear	Plum
5. Older Single	Lemon	Grape	Grape	Orange	Apple	Cherry	Pear	Pear	Plum
6. Older Couple	Lemon	Grape	Grape	Date	Apple	Cherry	Pear	Pear	Plum
7.Older family	Lemon	Grape	Grape	Apple	Apple	Cherry	Pear	Pear	Plum
8. Empty nester	Lemon	Grape	Grape	Date	Date	Cherry	Date	Pear	Plum
9.Ret'd couple	Lemon	Lemon	Grape	Date	Date	Date	Date	Pear	Plum
10. Ret'd single	Lemon	Lemon	Lemon	Date	Date	Date	Date	Pear	Pear

As in all science, it is important to ask the right question

“...it is clear that nothing limits ... the number of features according to which one can distribute [natural events or social facts] into several groups or distinct categories”

Cournot, 1843

Predicting classes: supervised classification

Huge number of tools

- linear discriminant analysis
- logistic regression
- tree classifiers
- nonparametric methods - e.g. nearest neighbour
- neural networks
- support vector machines
- .....

Very rich area
Data mining favours black box approach

Predicting measures: regression type models

Data mining tends to favour intensive use of simple methods over careful use of elaborate methods

So: multiple linear regression on subsets of variables and subsets of data instead of highly sophisticated model fitting investigation transformations of the variables, etc.

Of course, these are generalisations

Data visualisation

The human eye has evolved to detect shapes, structures, anomalies, and patterns:

Dynamic

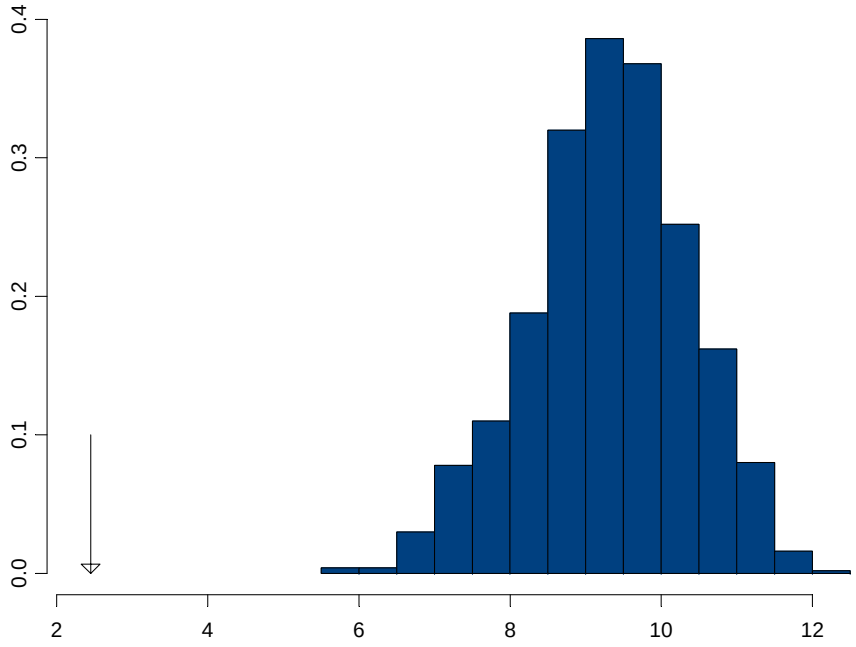
Interactive

Particularly useful for pattern discovery and systematic structures

Examples of pattern detection and discovery

Forelle and Bandler, *Wall Street Journal*, 2006

Anomaly detection in exam marks



New technical challenges

New Challenge Example 1: Streaming data

Definition: *streaming data* are data which arrive very rapidly, requiring analysis in real time

Examples: credit card fraud detection, speech recognition, algorithmic trading systems, etc

Cannot stop and analyse a batch of the data
Might not be able to store the data

Algorithms are based on summarising the data stream in some way and rolling only the summaries forward

Summaries could be
- a rolling window of data
- statistics: max to date, average to date, EWMA average, etc

Regular queries: know what summary stats to store
Ad hoc queries: not so easy

Housekeeping: how to compute the rolling statistics

e.g. to compute rolling variance

Compute rolling sums

$$S_1 = \sum^{n-1} x_t + x_n \qquad S_2 = \sum^{n-1} x_t^2 + x_n^2$$

and then the rolling variance as $S_2/n - \left(S_1/n\right)^2$

Inference:

e.g. 1: rounding inducing bias in variance estimate

Sample values independently from N(0,1)

Observe values rounded to nearest integer *x*:

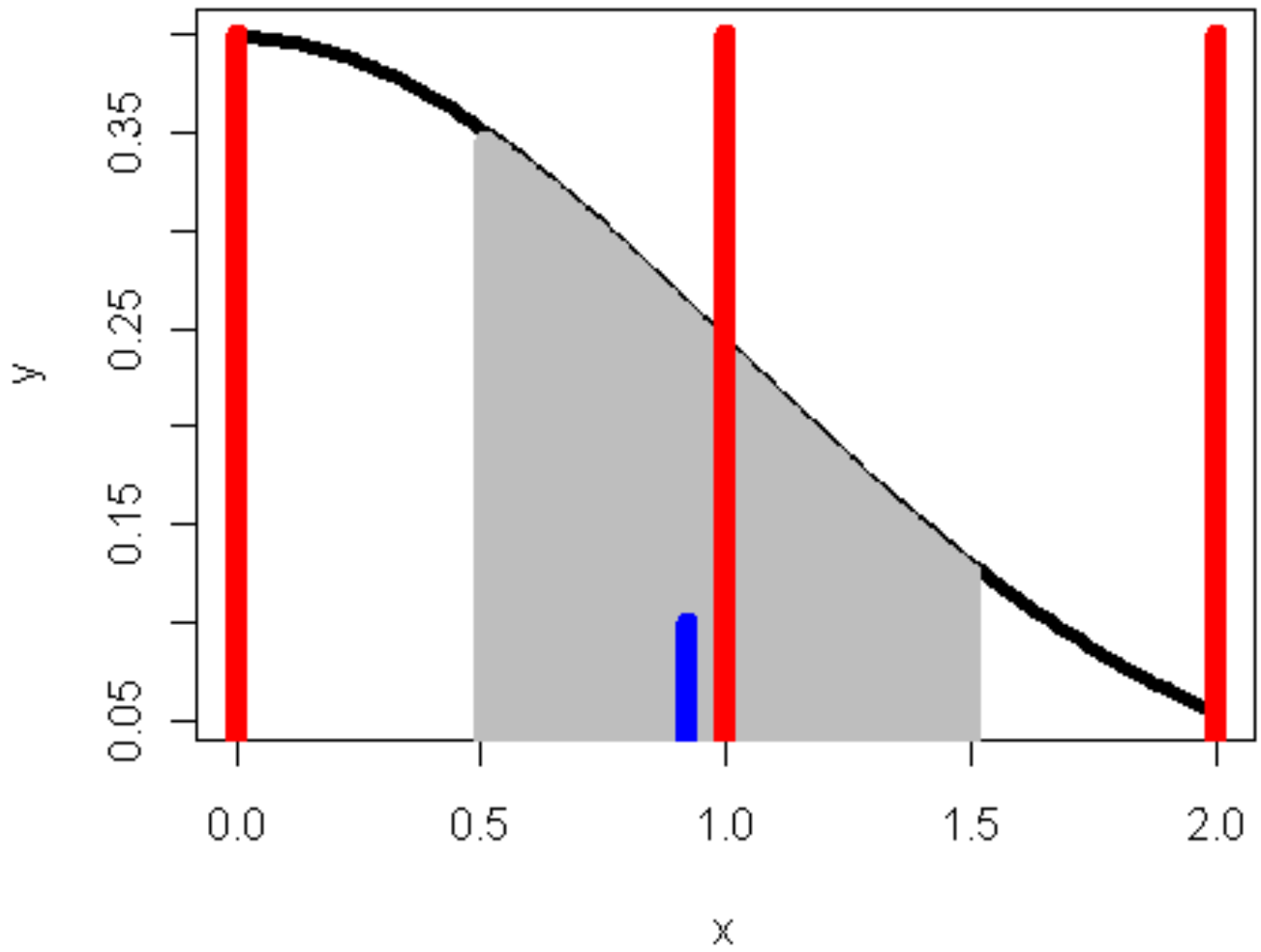
$r \in [x\text{-}0.4999... , x\text{+}0.4999...] \rightarrow x$

For r > 0 *there are more values between x*-0.4999 and *x* than there are between *x* and *x*+0.4999

That is, for *r* > 0 more observations are rounded *up* than *down*

and for *r* < 0 more observations are rounded *down* than *up*

so the estimated variance is biased up


y
0.05
0.15
0.25
0.35
0.0
0.5
1.0
1.5
2.0
x

e.g. 2: Rounding and autocorrelation in time series

Consider two consecutive values independently drawn from N(0,1): x1, x2

x1 + x2 is equally likely to be above or below x1

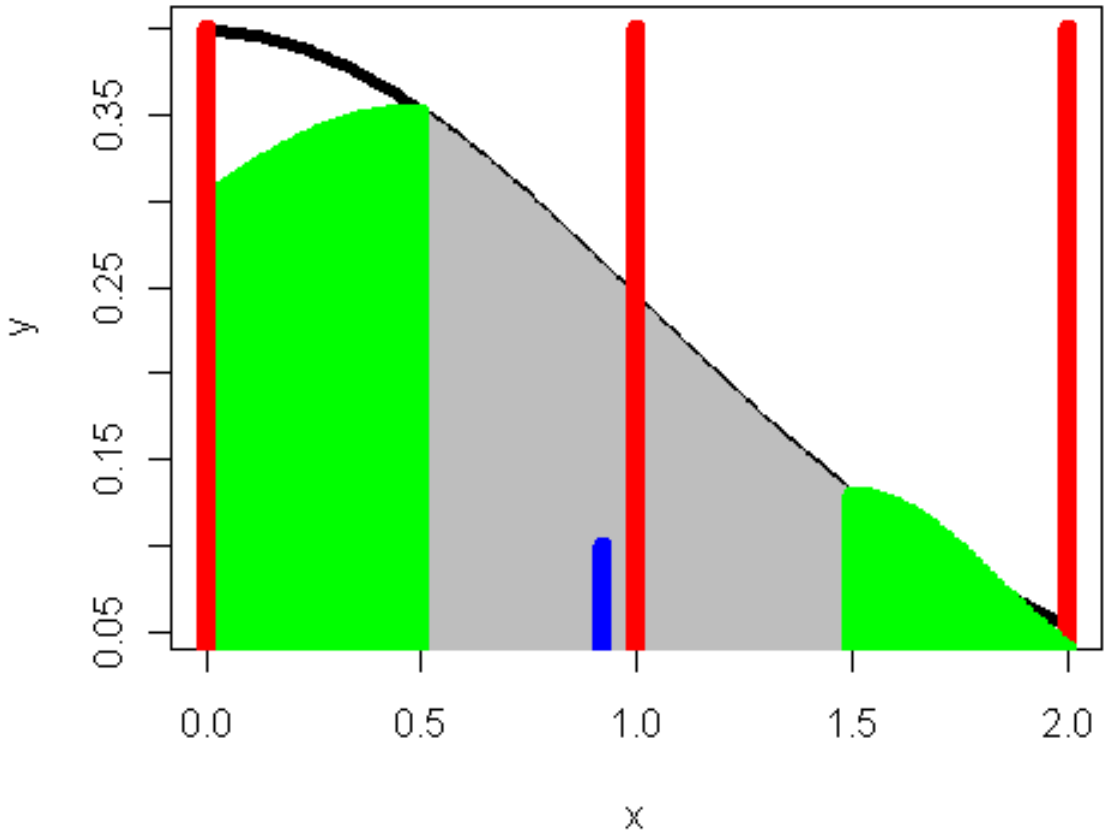
For the rounded data, x1r and (x1+x2)r, however ...

Suppose x1r = 1

This is more likely to have come from the interval [0.5, 1] than the interval [1, 1.5]. That is, it is more likely to have a true value less than its current value of 1

So (x1+x2) is more likely to have a value less than 0.5 than a value greater than 1.5

So (x1+x2)r is more likely to be 0 than 2


y
0.05
0.15
0.25
0.35
0.0
0.5
1.0
1.5
2.0
x

This is true whenever x1 > 0:
(x1+x2)r is more likely to be less than x1r than greater than x1r

And likewise, if x1 < 0
(x1+x2)r is more likely to be greater than x1r than less than x1r

For the original data
(x1 + x2) - (x1) is uncorrelated with x1

For the rounded data
(x1 + x2)r - x1r is negatively correlated with x1

So we have an induced negative correlation

e.g. 3: Length bias

I want to describe the typical person who visits a particular website

But computation means can only use a sample of $1/10^{th}$ of the visits

Solution?: randomly pick each visit with probability 1/10

But then those who visit more often are overrepresented

e.g. 4: suppose we want to know what proportion of queries were repeated, based on a sample of just 10% of the incoming stream

Include a query with probability /10

But: For a query that is made twice, its prob of being selected ***twice*** is

$$1/10 * 1/10 = 1/100$$

and its probability of being selected ***once*** is

$$2 * 1/10 * (9/10) = 18/100$$

So if s queries were made only once, and d queries were made twice, the proportion in our sample which were made twice is

$$(d/100) / [\, d/100 + [s/10 + 18d/100] \,] = d / [d + 10s + 18d]$$
$$= d / [\, 10s + 19d \,].$$

But the correct answer is d/[s+d]

New Challenge Example 2: Network data

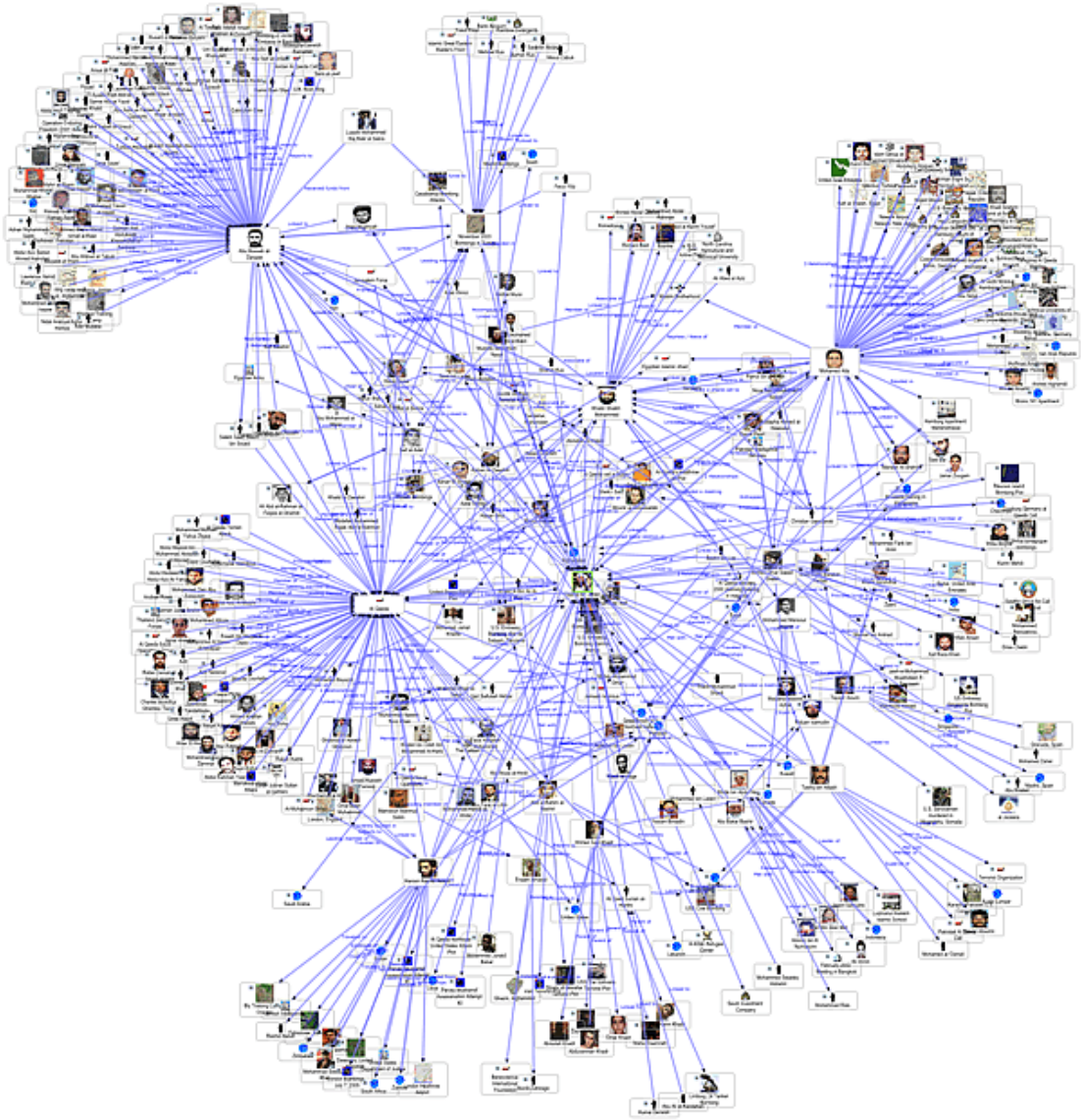
Link analysis: to evaluate connections between nodes of a graph

e.g. between people, organisations transactions

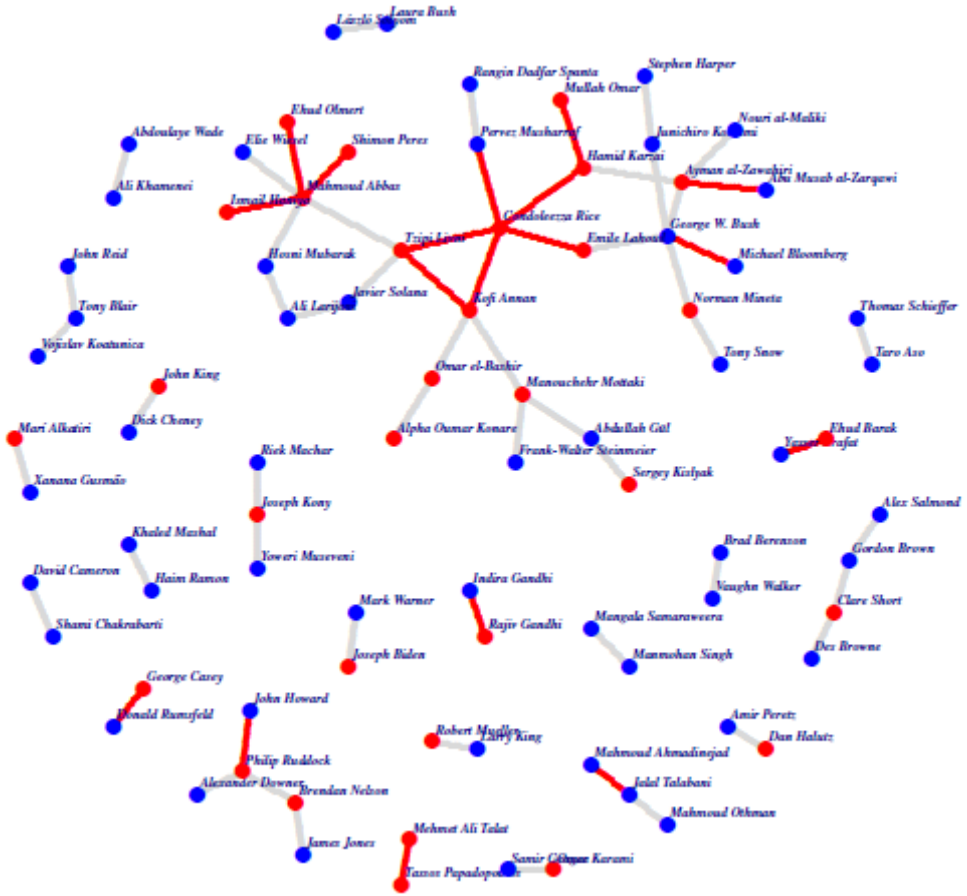
e.g. mortgage fraud rings, insider trading

e.g. scientific documents as nodes

Link analysis for Al-Quaeda terrorist network



All active individuals for week ending 28 June 2006, from European Media Monitor, anomalies in red



Many questions which might use network data:

- dynamic networks
- anomalies
- changes
- cliques
- key nodes
- outliers
- ....

New Challenge Example 3: Inference

Data mining involves massive searches, through many models, looking at many subsets of data

Raising inferential problems

Such as the familiar ***multiplicity*** problem

For a single test, the probability of incorrectly rejecting each true null hypothesis is controlled at α

But if we conduct two independent tests, and both null hypotheses are true

('independent': the outcome of one test is not predictable of or by the outcome of the other)

Then the probability of incorrectly rejecting *at least one* of the two true null hypotheses is

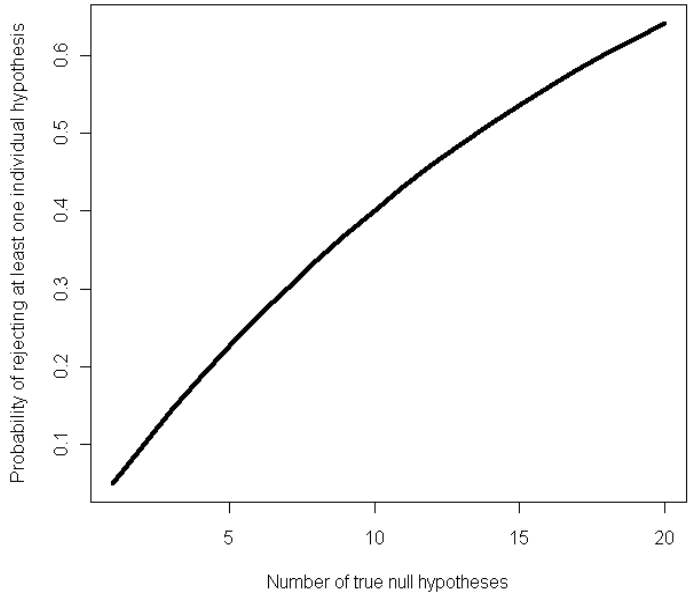
P(rej at least one) = 1 - P(not rejecting either)

$$= 1 - (1-\alpha)^2 \quad = \alpha(2-\alpha) \quad > \alpha$$

(last step since $\alpha < 1$)

e.g. $\alpha = 0.01 \Rightarrow$ P(rej at least one) = $\alpha(2-\alpha) = 0.01 \times 1.99 \approx 0.02$

If we conduct *N* independent tests, controlling each at the level α, then the probability of rejecting ***at least one*** of the null hypotheses, even if they are all true, is $1-(1-\alpha)^N$. $(1-\alpha)^N$ gets smaller as *N* increases, so $1-(1-\alpha)^N$ gets larger



e.g. if $\alpha = 0.05$ and $N = 100$, $1-(1-\alpha)^N = 0.99$

New perspectives

Closer look at exactly what the question is

		Decision		
		Accept	Rej	
True hypothesis	Null	*a*	*b* *I*	*T*
	Alternative	*c* *II*	*d*	*F*
		A	*R*	*N*

R is an observable random variable

1) *a*, *b*, *c*, *d* are unobservable

2) may be interested in various functions of *a*, *b*, *c*, *d*

A) Those based on b

(Let $F_b(b)$ be the distribution of b)

		Decision		
		Accept	Rej	
True hypothesis	Null	*a*	*b* *I*	*T*
	Alternative	*c* *II*	*d*	*F*
		A	*R*	*N*

A.1: Familywise (experimentwise) e.r., *FWER*: $P(b>0)=1-F_b(0)$

A.2: Generalised familywise e.r., *gFWER(k)*: $P(b>k)=1-F_b(k)$

A.3: Per comparison e.r. (or 'comparisonwise e.r.), *PCER*: $E(b)/N$

A.4: Per family e.r., *PFER*: $E(b)$

A.5: Median-based per family e.r., *mPFER*: $median(b)=F_b^{-1}(1/2)$

A.6: Quantile number of false positives, *QNFP*: $F_b^{-1}(\delta)$

B) Those based on b/R

[Let $F_{b/R}(b/R)$ be the distribution of b/R]

B.1: False discovery proportion (positive predictive value), *FDP*:

b/R (defined as 0 if R=0)

B.2: Tail probability for the proportion of false positives among the rejected hypotheses, *TPPFP*:

$$P(b/R > q) = 1 - F_{b/R}(q)$$

B.3: False discovery rate, *FDR*: $E(b/R)$

And Posterior error rate, *PER*: $P(b = 1 \mid R = 1)$

B.4: Positive false discovery rate, *pFDR*:

$$E(b/R \mid R > 0)$$

B.5: Proportion of expected false positives, *PEFP*:

$$E(b)/E(R)$$

B.6: Quantile proportion of false positives, *QPFP*:

$$F_{b/R}^{-1}(q)$$

“Every man should get to pick his own error rates.”

Miller (1981, p33)

The future

New data sources

new kinds of data

new opportunities

new problems

new challenges

Automatic data capture

Some people think that the advent of the internet (web searches, twitter, etc) has rendered traditional approaches to collecting social science data obsolete (surveys, administrative data sources, etc)

But need consistency for comparability
Beware of selection bias (e.g. self-selected surveys)

Do you think Twitter will be here in 100 years?
10 years?

Just one example:

Crime maps

- Originally trialled in Chicago
- Regularly updated maps showing the location and date of crimes
- *Police perspective*: enable better decisions, better targeting of resources, and improved tactics
- *Public's perspective*: enable citizens to identify risky areas to avoid, and to demand more police action if necessary
- From May of 2012, the UK public will also be able to see what action or outcome has occurred after a crime has been reported

But

Quality:

December 2011, Surrey Street, in Portsmouth was reported as having 136 crimes when it had actually had just two

Law of Unintended Consequences:

Survey by Direct Line Insurance found that 11% of respondents claim to have seen but not reported an incident ‘because they were scared it would drive away potential purchasers or renters’

Conclusion:

the discovery of interesting, unexpected, valuable structures in large data sets

- scalability

- increasing numbers of increasingly large data sets

- messy data; complicated data

- statistics (models) + comp.sci. (algorithms)

- global summaries vs local anomalies

Huge opportunities

With concomitant challenges

END

d.j.hand@imperial.ac.uk