Examples of large-scale network data

- World wide web (seemingly between 100 million and 1 billion websites)
- Social and communication networks (almost 300 million Twitter users)
- Computer networks and the “Internet of Things”
Research opportunities of Big Data

- Commercial/marketing benefits (funded by?)
- Understanding and influencing human society (iffy)
- Artificial intelligence (sci-fi?)
- Cyber-security
The cyber-security application

- Global cost of cyber-security is estimated at $400 billion (CSIC, 2014)
- Botnet behind a third of the spam sent in 2010: earned about $2.7 million
- Spam prevention: cost about > $1 billion (Anderson et al., 2013)
  - International terrorism affecting the UK or its interests, including a chemical, biological, radiological or nuclear attack by terrorists; and/or a significant increase in the levels of terrorism relating to Northern Ireland.
  - Hostile attacks upon UK cyber space by other states and large scale cyber crime.
  - A major accident or natural hazard which requires a national response, such as severe coastal flooding affecting three or more regions of the UK, or an influenza pandemic.
  - An international military crisis between states, drawing in the UK, and its allies as well as other states and non-state actors.
Statistics and cyber-security

Network flow data at Los Alamos National Laboratory: \(\sim 30\) GB/day

Typical attack pattern:

A. Opportunistic infection
B. Network traversal
C. Data exfiltration

Figure: Network traversal, source: Neil et al. (2013)
Posterior predictive p-value

The posterior predictive p-value is (Meng, 1994; Gelman et al., 1996, Eq. 2.8, Eq. 7)

\[ P = P\{f(D^*, \theta) \geq f(D, \theta) \mid D\}, \]

(1)

where \( \theta \) represents the model parameters, \( D \) is the observed dataset, \( D^* \) is a hypothetical replicated dataset generated from the model with parameters \( \theta \), and \( P(\cdot \mid D) \) is the joint posterior distribution of \((\theta, D^*)\) given \( D \).

In words: if a new dataset were generated from the same model and parameters, what is the probability that the new discrepancy would be as large?


Posterior sample: $\theta_1, \ldots, \theta_n$

Estimate I:
1. Simulate data $D_1^*, \ldots, D_n^*$.
2. Estimate
   $$\hat{P} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}\{f(D_i^*, \theta_i) \geq f(D, \theta_i)\}.$$ 

Estimate II:
1. Calculate p-values
   $$Q_i = P\{f(D_i^*, \theta_i) \geq f(D, \theta_i)\},$$
   for $i = 1, \ldots, n$.
2. Estimate
   $$\hat{P} = \frac{1}{n} \sum_{i=1}^{n} Q_i.$$
Estimation

Posterior sample: \( \theta_1, \ldots, \theta_n \)

Estimate I:

1. Simulate data \( D_1^*, \ldots, D_n^* \).
2. Estimate

\[
\hat{P} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{f(D_i^*, \theta_i) \geq f(D, \theta_i)\}.
\]

Estimate II:

1. Calculate p-values

\[
Q_i = P\{f(D^*, \theta_i) \geq f(D, \theta_i)\},
\]

for \( i = 1, \ldots, n \).
2. Estimate

\[
\hat{P} = \frac{1}{n} \sum_{i=1}^{n} Q_i.
\]
The two main points of this talk are:

1. Posterior predictive p-values are not just Bayesian model-checking devices; they can be used for hypothesis testing and anomaly detection.

2. Posterior predictive p-values are not uniformly distributed under the null hypothesis, but do come from a special family of distributions.
Recall:

\[ P = P \{ f(D^*, \theta) \geq f(D, \theta) \mid D \}. \]

How should we interpret \( P \)? For example:

- Is \( P = 40\% \) ‘good’?
- If \( P = 10^{-6} \), is there cause for alarm?
- What if 500 tests are performed, and \( \min(P_i) = 10^{-6} \)?

The problem is: \( P \) is not uniformly distributed.
Example: testing for dependence between point processes

A critical network question: does one point process ‘cause’ another? Possible applications:

- Characterizing information flow/network behaviour
- Detecting tunnelling on a computer network
- Finding ‘coordinated’ activity, e.g. discovering botnets
- Sports, finance, neurophysiology, etc..

One statement of the problem: Let $A$ and $B$ be two point processes and consider the hypothesis test

\[ H_0 : A \text{ and } B \text{ are independent} \]
\[ H_1 : B \text{ has an increased rate following } A \text{ events} \]
Example: testing for dependence between point processes

A critical network question: does one point process ‘cause’ another? Possible applications:

- Characterizing information flow/network behaviour
- Detecting tunnelling on a computer network
- Finding ‘coordinated’ activity, e.g. discovering botnets
- Sports, finance, neurophysiology, etc..

One statement of the problem: Let $A$ and $B$ be two point processes and consider the hypothesis test

$$H_0 : A \text{ and } B \text{ are independent}$$
$$H_1 : B \text{ has an increased rate following } A \text{ events}$$
Null hypothesis: $B$ is non-homogeneous Poisson process with intensity $r_0(t)$
Let $\tilde{b}_1$ = area until first response time
Let $\tilde{b}_2$ = area until second response time
Let $\tilde{b}_3 =$ area until third response time
Lemma

Under $H_0$, the areas $\tilde{b}_1, \tilde{b}_2, \ldots$ are the event times of a homogeneous Poisson process.
Alternative hypothesis: $B$ has intensity $r_1(t) = r_0(t)f(t - a(t))$, $f(x) \propto \exp(-\beta x)$ and $a(t)$ is closest $A$ event to $t$. 
Alternative hypothesis: \( B \) has intensity \( r_1(t) = r_0(t)f(t - a(t)), \) \( f(x) \propto \exp(-\beta x) \) and \( a(t) \) is closest \( A \) event to \( t. \)
Alternative hypothesis: $B$ has intensity $r_1(t) = r_0(t)f(t - a(t))$, $f(x) \propto \exp(-\beta x)$ and $a(t)$ is closest $A$ event to $t$
Alternative hypothesis: \( B \) has intensity \( r_1(t) = r_0(t)f(t - a(t)), \) \( f(x) \propto \exp(-\beta x) \) and \( a(t) \) is closest \( A \) event to \( t \).
Alternative hypothesis: $B$ has intensity $r_1(t) = r_0(t)f(t - a(t))$, $f(x) \propto \exp(-\beta x)$ and $a(t)$ is closest $A$ event to $t$
Alternative hypothesis: $B$ has intensity $r_1(t) = r_0(t)f(t - a(t))$, where $f$ is a step-function with one step
Alternative hypothesis: $B$ has intensity $r_1(t) = r_0(t)f(t - a(t))$, where $f$ is a step-function with one step
Alternative hypothesis: $B$ has intensity $r_1(t) = r_0(t)f(t - a(t))$, where $f$ a decreasing function. ($\hat{f}$ is maximum likelihood)
Alternative hypothesis: $B$ has intensity $r_1(t) = r_0(t)f(t - a(t))$, where $f$ a decreasing function. ($\hat{r}_1$ is maximum likelihood)
Theorem

Under $H_0$, $\hat{f}(0)$ is distributed as $n/(UT)$ where $U$ is a uniform random variable over $[0, 1]$ and $T$ is the length of the observation period.

p-value for $A$ causing $B$: $n/\{T\hat{f}(0)\}$

In previous example: $p \approx 0.15$ (not significant)
Daily emailing behaviour of an individual in the Enron dataset:

![Graph showing daily emailing behaviour over time and days. The x-axis represents time from 0 to 24 hours, and the y-axis represents days from 0 to 365 days. The graph is a scatter plot with numerous data points.](image-url)
Bayesian intensity estimation for an individual of interest (change-point model for daily binned data, and density estimation for within day behaviour)
‘Information flow’ through JD in the Enron data. Over 2001, 12 individuals contact and are contacted back by JD. Posterior predictive p-values are calculated for each $i \rightarrow JD \not\Rightarrow JD \rightarrow k$, $i, k = 1, \ldots, 12$. Full black $p < 0.0001$, half-black means $p \leq 0.05$, white means not significant.
$7 \to j \not\leftrightarrow j \to 7$: only one email on each edge, they are sent about one month apart, and appear to be unrelated judging by their subject-lines. The $p$-value is computed to be about .2

$10 \to j \not\leftrightarrow j \to 10$: 14 emails from 10 to $i$ and 9 from $i$ to 10, the most coincidental email times falling in July, about $3\frac{1}{2}$ hours from each other. The $p$-value is 0.07. If time is not transformed the raw $p$-value is 0.0035, suggesting a significant interaction. However, upon inspecting the subject-lines of $10 \to i$ and $i \to 10$ it is not in fact obvious that there is consistent reciprocation. For example, the subject-lines of the two most coincidental emails are “FW: Enron Complaint” and “Dunn hearing link?”, which are not obviously related.
Based on the subject-lines of the most coincidental email times, the three black circles are correct detections.

6 $\rightarrow i \leadsto i \rightarrow 5$ The $p$-value is around $2 \cdot 10^{-5}$. The most coincidental email times are two hours and twenty minutes apart, and have the same subject-line, “Re: FW: SoCalGas Capacity”.

9 $\rightarrow i \leadsto i \rightarrow 3$ The $p$-value is around $6 \cdot 10^{-8}$. The most coincidental email times are 50 minutes apart, have the subject-lines “California Update–Legislative Push Underway” and “Re: California Update–Legislative Push Underway”.

12 $\rightarrow i \leadsto i \rightarrow 11$ The $p$-value is around $2 \cdot 10^{-5}$. The most coincidental email times are 30 minutes apart and have the subject-lines “RE: CA Unbundling” and “CA Unbundling”
The notion that PPPs are ‘conservative’

General consensus is that $P$ is conservative (too large), i.e. under-represents evidence against the model, see e.g. discussion of Gelman et al. (1996).

In fact, conservative isn’t quite the right word. Instead, the p-values seem to concentrate around $1/2$. For example, let $D \sim \text{normal}(\mu, 1)$, with a vague prior on $\mu$ and let $f(D, \theta) = D$. Then,

$$P = \Pr\{f(D^*, \theta) \geq f(D, \theta) \mid D\} \approx 1/2.$$
Meng’s result

In the last pages of Meng (1994) we find:

**Theorem 1.** Suppose, given \( \psi \in \Psi_0 \), the sampling distribution of a discrepancy measure \( D(X, \psi) \) is continuous. Then under the prior predictive distribution (5.2), the corresponding posterior predictive p-value, \( p_B \), is stochastically less variable than a uniform distribution but with the same mean; that is, if \( U \) is uniformly distributed on \([0, 1]\), then (i) \( E(p_B) = E(U) = \frac{1}{2} \) and (ii) \( E[h(p_B)] \leq E[h(U)] \) for all convex functions \( h \) on \([0, 1]\), where the expectations involving \( p_B \) are with respect to (5.2).

In our notation: suppose \( \theta \) and \( D \) are drawn from the prior and model respectively, and \( f(D, \theta) \) is absolutely continuous. Then, (i) \( E(P) = E(U) = 1/2 \) and (ii) for any convex function \( h \),

\[
E\{h(P)\} \leq E\{h(U)\},
\]

where \( U \) is a uniform random variable on \([0, 1]\).
Meng (1994) next finds:

**Lemma 1.** Let $G(\alpha)$ be the c.d.f. of a random variable $W$ on $[0, 1]$. If $W$ is stochastically less variable than $U[0, 1]$ but with the same mean, then $\forall \alpha \in [0, 1]$

\[
\alpha - \left[ \alpha^2 - 2 \int_0^\alpha G(t) \, dt \right]^{1/2} \leq G(\alpha) \leq \alpha + \left[ \alpha^2 - 2 \int_0^\alpha G(t) \, dt \right]^{1/2} \leq 1.
\]

The first or second inequality becomes an equality for all $\alpha$ if and only if $G(\alpha) \equiv \alpha$.

One direct consequence of (5.5) is that

\[
G(\alpha) \leq 2\alpha \quad \text{for all } \alpha \leq \frac{1}{2}.
\]
The convex order

Let $X$ and $Y$ be two random variables with probability measures $\mu$ and $\nu$ respectively. We say that $\mu$ (respectively, $X$) is less variable than $\nu$ (respectively, $Y$) in the convex order, denoted $\mu \leq_{cx} \nu$ (or $X \leq_{cx} Y$) if, for any convex function $h$, $\mathbb{E}\{h(X)\} \leq \mathbb{E}\{h(Y)\}$, whenever the expectations exist (Shaked and Shanthikumar, 2007). A probability measure $\mathcal{P}$, and a random variable distributed as $\mathcal{P}$, is sub-uniform if $\mathcal{P} \leq_{cx} \mathcal{U}$, where $\mathcal{U}$ is a uniform distribution on $[0, 1]$.

Posterior predictive p-values have a sub-uniform distribution.
Meng’s findings could seem quite coarse:

1. The $2\alpha$ bound suggests that $P$ could be liberal, whereas we had the impression they were conservative. Can the bound be improved?

2. The theorem seems to be describing a huge space of distributions. Can it somehow be reduced?

Our results:

1. It is possible to construct a posterior predictive p-value with any sub-uniform distribution.

2. Some sub-uniform distributions achieve the $2\alpha$ bound.

3. Therefore, some posterior predictive p-values achieve the $2\alpha$ bound. In fact, we can construct simple examples where this happens.
Meng’s findings could seem quite coarse:

1. The $2\alpha$ bound suggests that $P$ could be liberal, whereas we had the impression they were conservative. Can the bound be improved?
2. The theorem seems to be describing a huge space of distributions. Can it somehow be reduced?

Our results:

1. It is possible to construct a posterior predictive p-value with any sub-uniform distribution.
2. Some sub-uniform distributions achieve the $2\alpha$ bound.
3. Therefore, some posterior predictive p-values achieve the $2\alpha$ bound. In fact, we can construct simple examples where this happens.
Theorem (Strassen’s theorem)

For two probability measures $\mu$ and $\nu$ on the real line the following conditions are equivalent:

1. $\mu \leq_{cx} \nu$;
2. there are random variables $X$ and $Y$ with marginal distributions $\mu$ and $\nu$ respectively such that $E(Y \mid X) = X$.

Theorem due to Strassen (1965) (see also references therein), this version due to Müller and Rüschendorf (2001).
Let \( \mu \) and \( \nu \) be two probability measures on the real line where \( \nu \) is absolutely continuous. The following conditions are equivalent:

1. \( \mu \preceq_{cx} \nu \);
2. there exist random variables \( X \) and \( Y \) with marginal distributions \( \mu \) and \( \nu \) respectively such that \( E(Y \mid X) = X \) and the random variable \( Y \mid X \) is either singular, i.e. \( Y = X \), or absolutely continuous with \( \mu \)-probability one.

**Theorem (Posterior predictive p-values and the convex order)**

\( \mathcal{P} \) is a sub-uniform probability measure if and only if there exist random variables \( \mathcal{P}, D, \theta \) and an absolutely continuous discrepancy \( f(D, \theta) \) such that

\[
P = \mathcal{P}\{f(D^*, \theta) \geq f(D, \theta) \mid D\},
\]

where \( \mathcal{P} \) has measure \( \mathcal{P} \), \( D^* \) is a replicate of \( D \) conditional on \( \theta \) and \( \mathcal{P}(\cdot \mid D) \) is the joint posterior distribution of \( (\theta, D^*) \) given \( D \).
Theorem

Let \( \mu \) and \( \nu \) be two probability measures on the real line where \( \nu \) is absolutely continuous. The following conditions are equivalent:

1. \( \mu \leq_{cx} \nu \);
2. there exist random variables \( X \) and \( Y \) with marginal distributions \( \mu \) and \( \nu \) respectively such that \( E(Y \mid X) = X \) and the random variable \( Y \mid X \) is either singular, i.e. \( Y = X \), or absolutely continuous with \( \mu \)-probability one.

Theorem (Posterior predictive p-values and the convex order)

\( \mathcal{P} \) is a sub-uniform probability measure if and only if there exist random variables \( P, D, \theta \) and an absolutely continuous discrepancy \( f(D, \theta) \) such that

\[
P = P\{f(D^*, \theta) \geq f(D, \theta) \mid D\},
\]

where \( P \) has measure \( \mathcal{P} \), \( D^* \) is a replicate of \( D \) conditional on \( \theta \) and \( P(\cdot \mid D) \) is the joint posterior distribution of \( (\theta, D^*) \) given \( D \).
Exploring the set of sub-uniform distributions

The \textit{integrated} distribution function of a random variable $X$ with distribution function $F_X$ is,

$$\psi_X(x) = \int_{-\infty}^{x} F_X(t) dt.$$ 

We have (Müller and Rüschendorf, 2001)

1. $\psi_X$ is non-decreasing and convex;
2. Its right derivative $\psi_X^+(x)$ exists and $0 \leq \psi_X^+(x) \leq 1$;
3. $\lim_{x \to -\infty} \psi_X(x) = 0$ and $\lim_{x \to \infty} \{x - \psi_X(x)\} = E(X)$.

Furthermore, for any function $\psi$ satisfying these properties, there is a random variable $X$ such that $\psi$ is the integrated distribution function of $X$. The right derivative of $\psi$ is the distribution function of $X$, $F_X(x) = \psi^+(x)$.

Let $Y$ be another random variable with integrated distribution function $\psi_Y$. Then $X \leq_{cx} Y$ if and only if $\psi_X \leq \psi_Y$ and $\lim_{x \to \infty} \{\psi_Y(x) - \psi_X(x)\} = 0$. 

Exploring the set of sub-uniform distributions

The *integrated* distribution function of a random variable $X$ with distribution function $F_X$ is,

$$
\psi_X(x) = \int_{-\infty}^{x} F_X(t) \, dt.
$$

We have (Müller and Rüschendorf, 2001)

1. $\psi_X$ is non-decreasing and convex;
2. Its right derivative $\psi_X^+(x)$ exists and $0 \leq \psi_X^+(x) \leq 1$;
3. $\lim_{x \to -\infty} \psi_X(x) = 0$ and $\lim_{x \to \infty} \{x - \psi_X(x)\} = E(X)$.

Furthermore, for any function $\psi$ satisfying these properties, there is a random variable $X$ such that $\psi$ is the integrated distribution function of $X$. The right derivative of $\psi$ is the distribution function of $X$, $F_X(x) = \psi^+(x)$.

Let $Y$ be another random variable with integrated distribution function $\psi_Y$. Then $X \preceq_{cx} Y$ if and only if $\psi_X \preceq \psi_Y$ and $\lim_{x \to \infty} \{\psi_Y(x) - \psi_X(x)\} = 0$. 
Exploring the set of sub-uniform distributions

The *integrated* distribution function of a random variable $X$ with distribution function $F_X$ is,

$$
\psi_X(x) = \int_{-\infty}^{x} F_X(t) \, dt.
$$

We have (Müller and Rüschendorf, 2001)

1. $\psi_X$ is non-decreasing and convex;
2. Its right derivative $\psi_X^+(x)$ exists and $0 \leq \psi_X^+(x) \leq 1$;
3. $\lim_{x \to -\infty} \psi_X(x) = 0$ and $\lim_{x \to \infty} \{x - \psi_X(x)\} = E(X)$.

Furthermore, for any function $\psi$ satisfying these properties, there is a random variable $X$ such that $\psi$ is the integrated distribution function of $X$. The right derivative of $\psi$ is the distribution function of $X$, $F_X(x) = \psi^+(x)$.

Let $Y$ be another random variable with integrated distribution function $\psi_Y$. Then $X \leq_{cx} Y$ if and only if $\psi_X \leq \psi_Y$ and $\lim_{x \to \infty} \{\psi_Y(x) - \psi_X(x)\} = 0$. 

P. Rubin-Delanchy  
Bayesian anomaly detection for networks and Big Data
Exploring the set of sub-uniform distributions

The *integrated* distribution function of a random variable $X$ with distribution function $F_X$ is,

$$\psi_X(x) = \int_{-\infty}^{x} F_X(t) dt.$$ 

We have (Müller and Rüschendorf, 2001)

1. $\psi_X$ is non-decreasing and convex;
2. Its right derivative $\psi_X^+(x)$ exists and $0 \leq \psi_X^+(x) \leq 1$;
3. $\lim_{x \to -\infty} \psi_X(x) = 0$ and $\lim_{x \to \infty} \{x - \psi_X(x)\} = E(X)$.

Furthermore, for any function $\psi$ satisfying these properties, there is a random variable $X$ such that $\psi$ is the integrated distribution function of $X$. The right derivative of $\psi$ is the distribution function of $X$, $F_X(x) = \psi^+(x)$.

Let $Y$ be another random variable with integrated distribution function $\psi_Y$. Then $X \preceq_{cx} Y$ if and only if $\psi_X \leq \psi_Y$ and $\lim_{x \to \infty} \{\psi_Y(x) - \psi_X(x)\} = 0$. 

P. Rubin-Delanchy  Bayesian anomaly detection for networks and Big Data
Figure: Uniform distribution
Figure: Some sub-uniform distributions. Blue: uniform, green: Beta(2,2)
Figure: Some sub-uniform distributions. Blue: uniform, green: Beta(2,2), red: a distribution with maximal mass at .3.
Figure: Some constructive examples
Suppose $P_1, \ldots, P_m$ are independent posterior predictive p-values. How can I combine them into one piece of evidence?

With ordinary p-values (i.e. independent uniform random variables):

- $\min(P_i) \sim \text{Beta}(1, m)$
- $-2 \sum \log(P_i) \sim \chi^2_{2m}$ (Fisher’s method)
Lemma

Let $P_1, \ldots, P_m$ and $U_1, \ldots, U_m$ each be sequences of independent sub-uniform and uniform random variables on $[0, 1]$, respectively. For $x \in [0, 1]$, let

\[ q = 1 - (1 - x)^m = P\{\min(U_i) \leq x\}. \]

Then

\[ P\{\min(P_i) \leq x\} \leq 1 - (1 - 2x)^m, \]

\[ = 1 - \{2(1 - q)^{1/m} - 1\}^m, \]

which is no larger than $2q$ and tends to $2q - q^2$ as $m \to \infty$. Furthermore, this bound is achievable if the $P_i$ are independent and identically distributed.
Lemma (Fisher’s method is asymptotically conservative)

Let $P_1, \ldots, P_m$ and $U_1, \ldots, U_m$ each be sequences of independent and identically distributed sub-uniform and uniform random variables on $[0, 1]$ respectively. For $\alpha \in (0, 1]$, let $t_{\alpha,m}$ be the critical value defined by

$$P \left( -2 \sum_{i=1}^{m} \log(U_i) \geq t_{\alpha,m} \right) = \alpha.$$ 

Then there exists $n \in \mathbb{N}$ such that

$$P \left( -2 \sum_{i=1}^{m} \log(P_i) \geq t_{\alpha,m} \right) \leq \alpha,$$

for any $m \geq n$. 
Lemma (Sums of posterior predictive p-values)

Let $P_1, \ldots, P_m$ denote $m$ independent sub-uniform random variables with mean $\bar{P} = m^{-1} \sum_{i=1}^{m} P_i$. Then, for $0 \leq t \leq 1/2$, 

$$P \left( \frac{1}{2} - \bar{P} \geq t \right) \leq \exp(-6mt^2).$$

Note Hoeffing’s inequality (Hoeffding, 1963)

$$P \left( \frac{1}{2} - \bar{P} \geq t \right) \leq \exp(-2mt^2).$$
Some final points

1. Other attempts at combining Bayesian inference with hypothesis tests completely break down when the posterior contracts.

2. We can also characterize *estimated* posterior predictive p-values from MCMC samples, e.g. of two schemes proposed by Gelman et al. (1996), only one produces a sub-uniform estimate, and other types of p-values (e.g. discrete p-values).

3. The convex order seems to have a deeper relevance to statistical inference and Monte Carlo, see e.g. Andrieu and Vihola (2014), Goldstein et al. (2011)