A Deambular pela Internet...

M. Rosário de Oliveira e C. Pascoal

Departamento de Matemática, IST e CEMAT
Seminário de Matemática, LMAC

November 9, 2010

Sumário

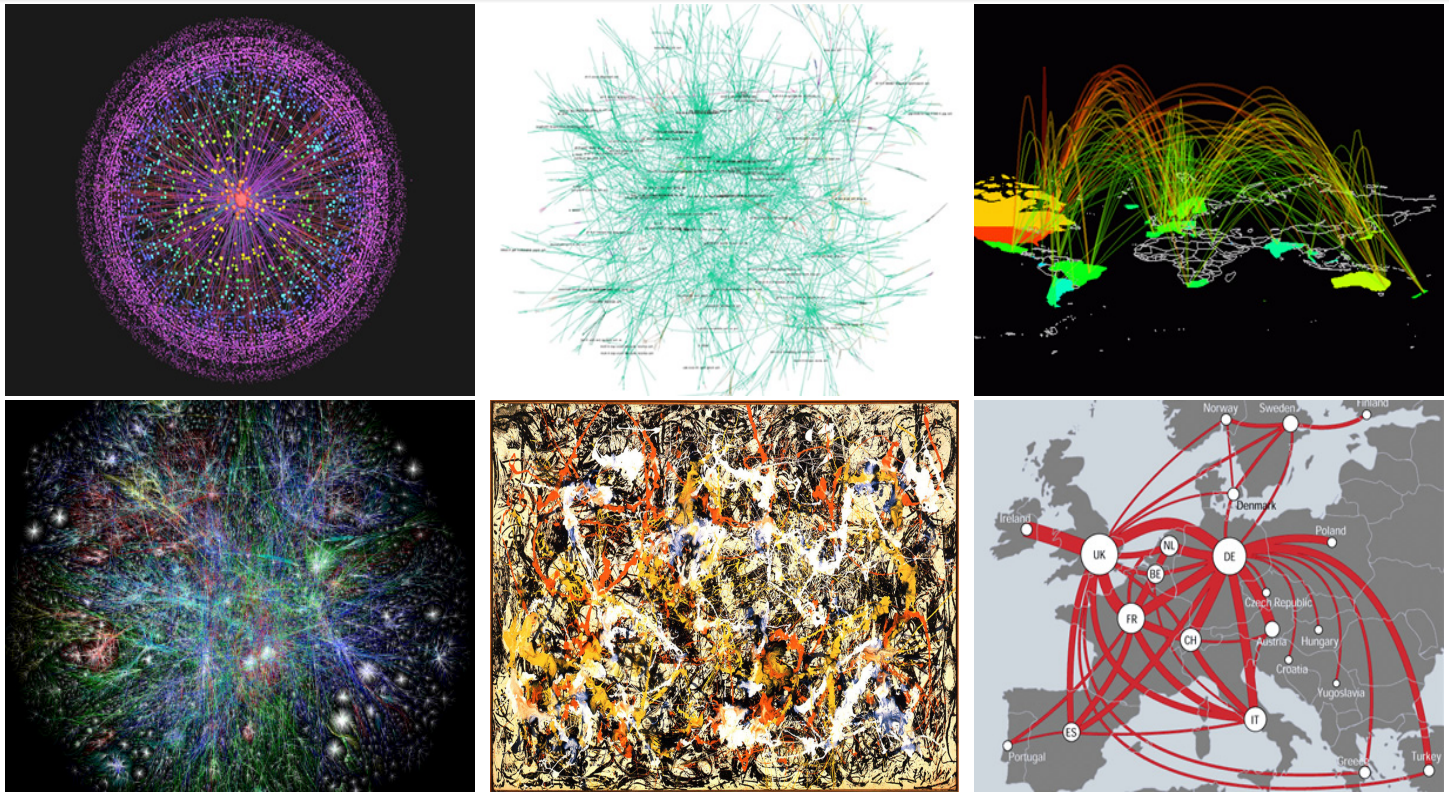
O que é a Internet?

- Sistema global de **redes de computadores** interligadas
- É uma **rede de redes** privadas, públicas, académicas, empresariais, governamentais...
- **Redefiniu** o que se entende por telefone, música, cinema, televisão...
- Permite novas formas de **interações entre as pessoas** através de SMS, fóruns da discussão e redes sociais
- Alargou possibilidades de **negócio**: compras online

Origem

- Leonard Kleinrock (1934-) é considerado o **pai da Internet**. Publicou em 1961 o primeiro trabalho sobre redes com comutação de pacotes. Era então estudante do MIT
- A **Web** foi inventada por Tim Berners-Lee (1955-) no CERN; desenvolveu de uma só vez o HTML (linguagem), o HTPP (protocolo) e as URL (endereços)
- O **primeiro servidor** Web foi colocado on-line pela primeira vez em 6 de Agosto de 1991
- A partir de 2009, um **quarto** da população estimada da Terra usou a Internet

Representações gráficas da Internet



- www.visualcomplexity.com & Convergence, Pollock

Motivação

Exemplos:

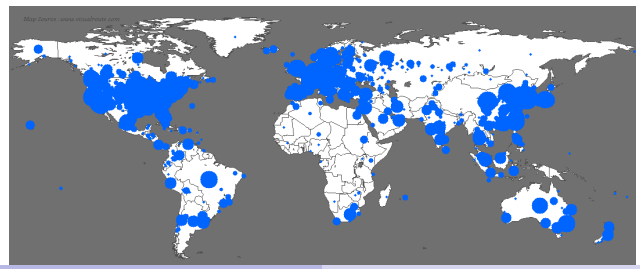
- **Detecção de fraudes** em cartões de crédito
- Distúrbios em Ecosistemas: **furacões, secas, inundações, incêndios**
- **Saúde**: imagem de ressonância magnética anómala pode indicar a presença de lesões malignas ou tumores
- **Vigilância militar** das actividades do inimigo

Causas Possíveis:

- Observações pertencendo a diferentes **classes**
- **Erros de medição**
- Presença de **ruído** nos dados
- . . .

O vírus *Sapphire/Slammer*

- O vírus começou a infectar máquinas um pouco antes 05:30 UTC num **Sábado,** 25 **de Janeiro de** 2003
- O **Vírus Sapphire** foi o vírus mais **rápido** a propagar-se
- Através da Internet, **duplicou** em tamanho em cada 8.5 segundos
- Infectou mais de 90% de máquinas vulneráveis em 10 **minutos**
- **Consequências:**
  - Pelo menos 75 000 máquinas **infectadas**
  - **Voos** foram **cancelados**
  - **Interferências** em **eleições**
  - **Transferências** a partir de máquinas multibanco **falharam**



Problema: Detectar anomalias no tráfego da Internet *(in real time!)*

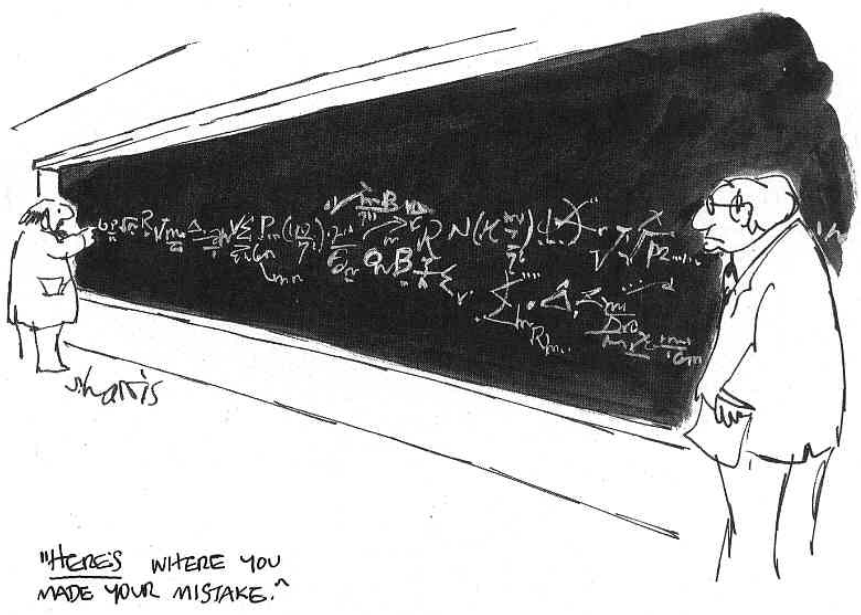
O que é uma anomalias no tráfego da Internet?

- Padrões **discordantes** da generalidade dos dados ou **atípicos**

Como resolver este problema?

ENGENHEIROS

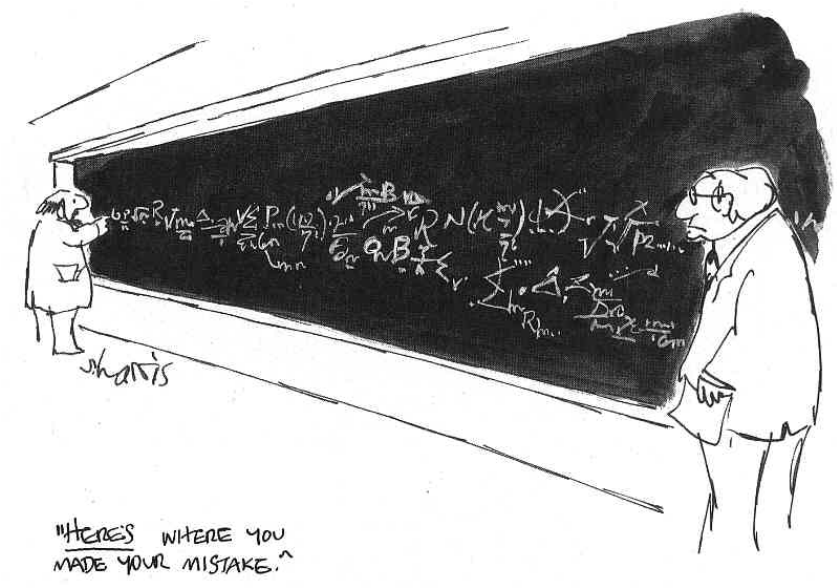
MATEMÁTICOS



ENGENHEIROS

- Preocupados com o problema de classificação

MATEMÁTICOS

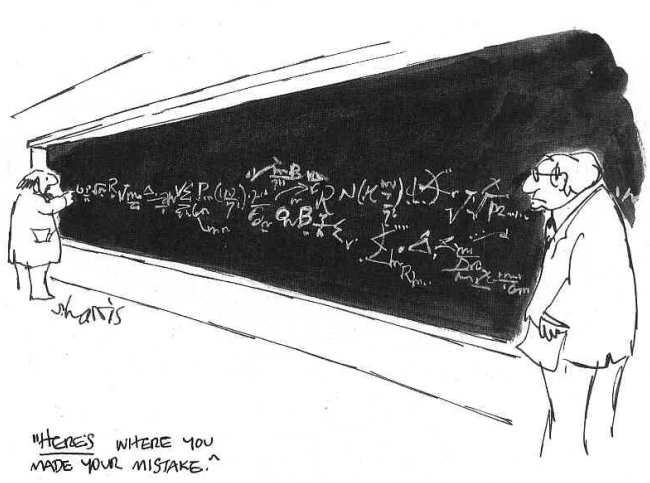


- Preocupados com o problema de estimação

ENGENHEIROS

- Preocupados com o problema de classificação
- Serão todas as anomalias detectadas? (recall)

MATEMÁTICOS

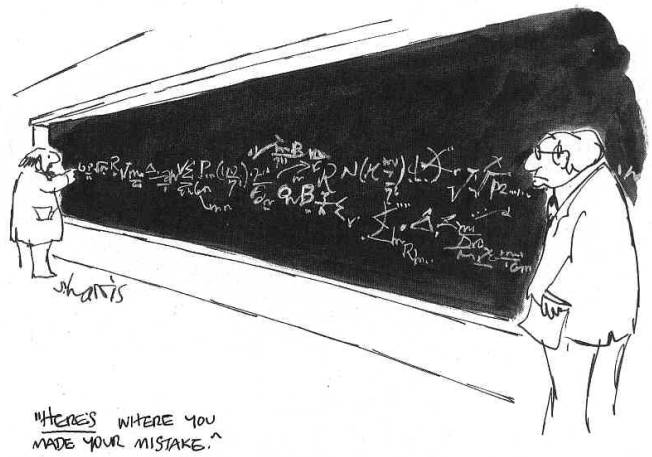


- Preocupados com o problema de estimação
- Os parâmetros são bem estimados?

ENGENHEIROS

- Preocupados com o problema de classificação
- Serão todas as anomalias detectadas? (recall)
- Entre os classificados como anomalias, existem fluxos regulares? (precision)

MATEMÁTICOS

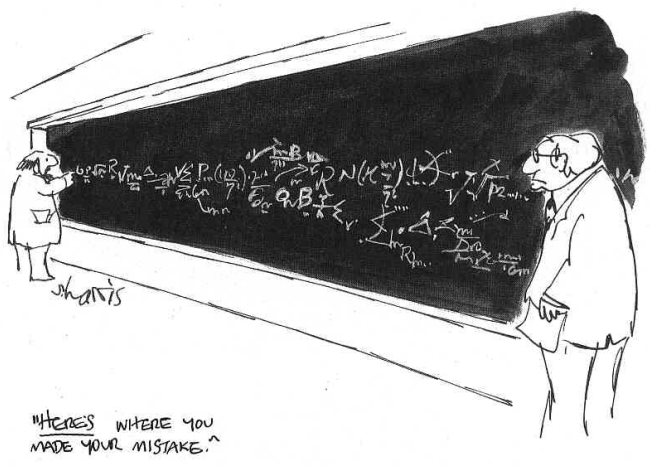


- Preocupados com o problema de estimação
- Os parâmetros são bem estimados?
- Propriedades dos estimadores

ENGENHEIROS

- Preocupados com o problema de classificação
- Serão todas as anomalias detectadas? (recall)
- Entre os classificados como anomalias, existem fluxos regulares? (precision)
- . . .

MATEMÁTICOS



- Preocupados com o problema de estimação
- Os parâmetros são bem estimados?
- Propriedades dos estimadores
- . . .

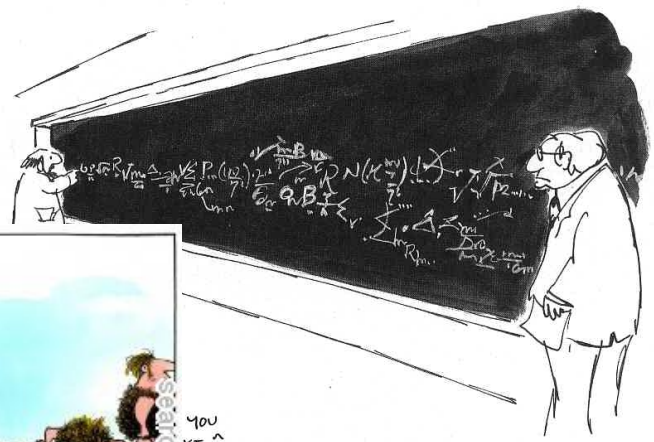
ENGENHEIROS

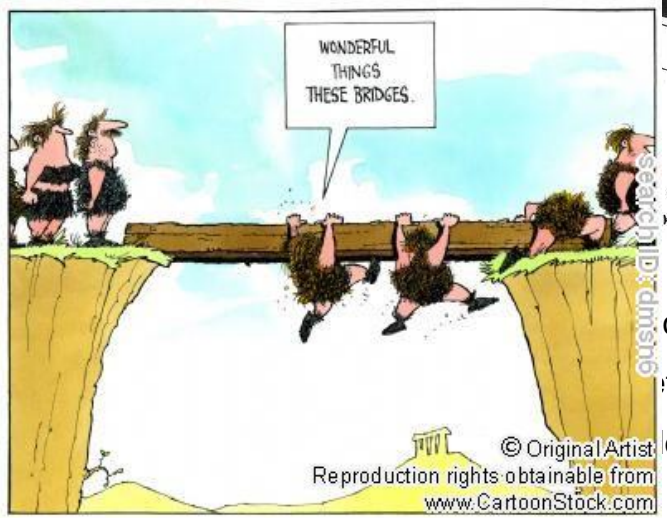
- Preocupados com o problema
- Serão todas as anomalias dete
- Entre os classificados como ar

  fluxos regulares? (precision)

- . . .

MATEMÁTICOS



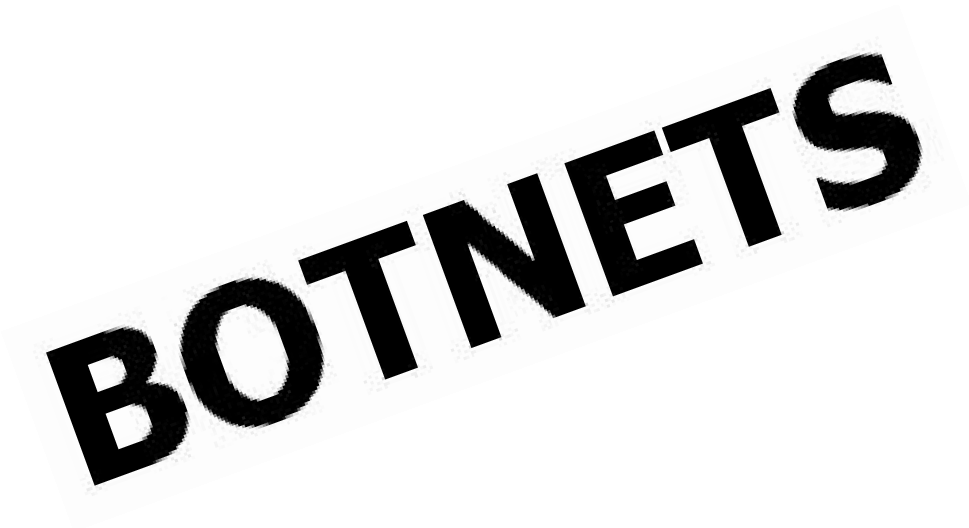


- os com o problema de estimação
- tros são bem estimados?
- es dos estimadores
- . . .

Dados emulados (Synthetic data)

A natureza das anomalias muda constantemente e os intrusos adaptam os seus ataques de modo a fugir às soluções de detecção de anomalias conhecidas!

Botnets



Dados emulados (Synthetic data)

- Dados sintéticos foram **emulados** numa **rede controlada**, pelo **Instituto de Telecomunicações** - Aveiro

- **Regulares:** **HTTP** + **Streaming** + **BitTorrent**
- **Anomalias:** **Snapshots** + **Nmaps**

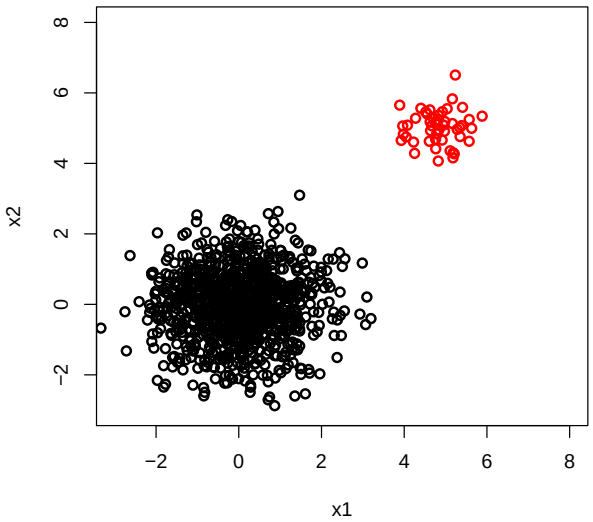
Variáveis ($p = 11$):

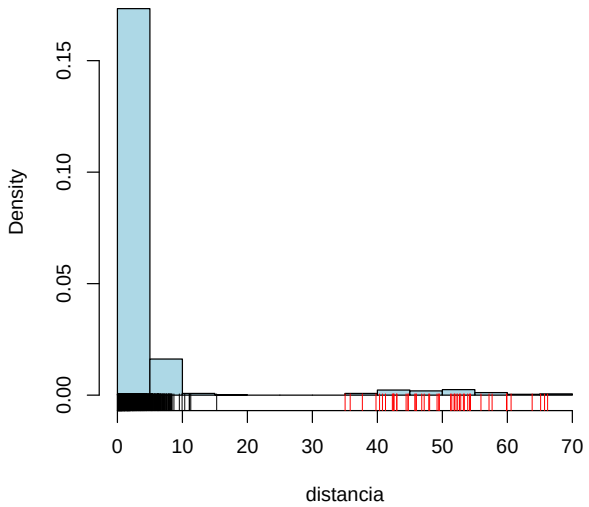
- **Bytes**: Média, Desvio-padrão (Up and Down)
- **Pacotes**: Média, Desvio-padrão (Up and Down)
- **Sessões**: Média, Desvio-padrão
- Estatística de **Fisher**: Número de Sessões

Procedimento

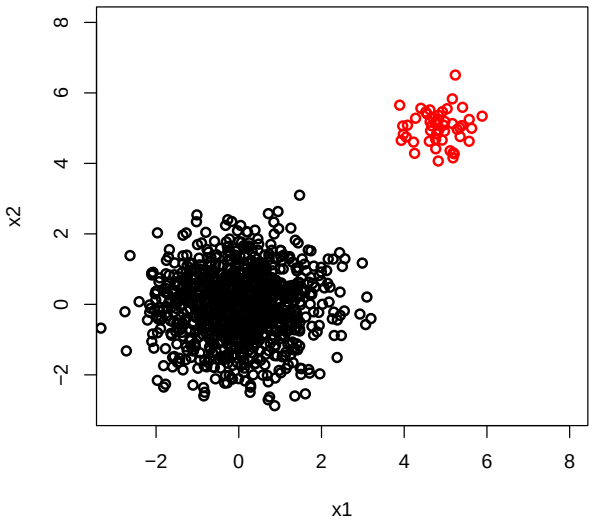
- Defina uma **transformação linear** os dados, de **menor dimensão**, tal que no novo espaço se evidenciam as diferenças entre tráfego anómalo e regular
- As observações **regulares**, quando projectadas neste espaço, têm um **padrão elíptico**
- As observações **anómalas**, projectadas neste espaço, são **extremas**
- Defina uma métrica e calcule a **distância** de cada observação ao **centro** dos dados
- As observações **mais distantes** do centro são consideradas **outliers** (anomalias)

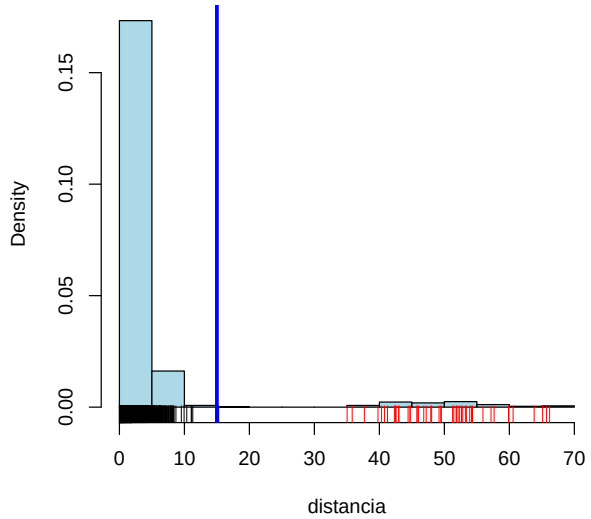
Procedimento





Procedimento



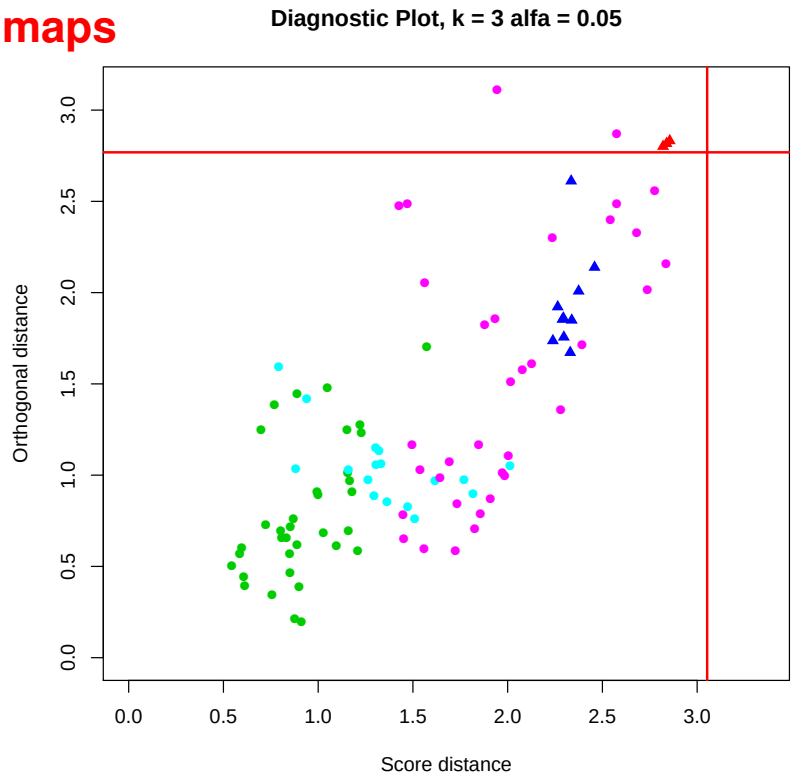


USA university

Observações ($n = 100$):

- **Regulares:** 35 **http** + 17 **streaming** + 35 **BitTorrent**
- **Contaminação:** 10 **snapshots** + 3 **nmaps**
- $k = 3$

	PCAGRID
False Positive	0.023
Recall	0.231
Precision	0.600



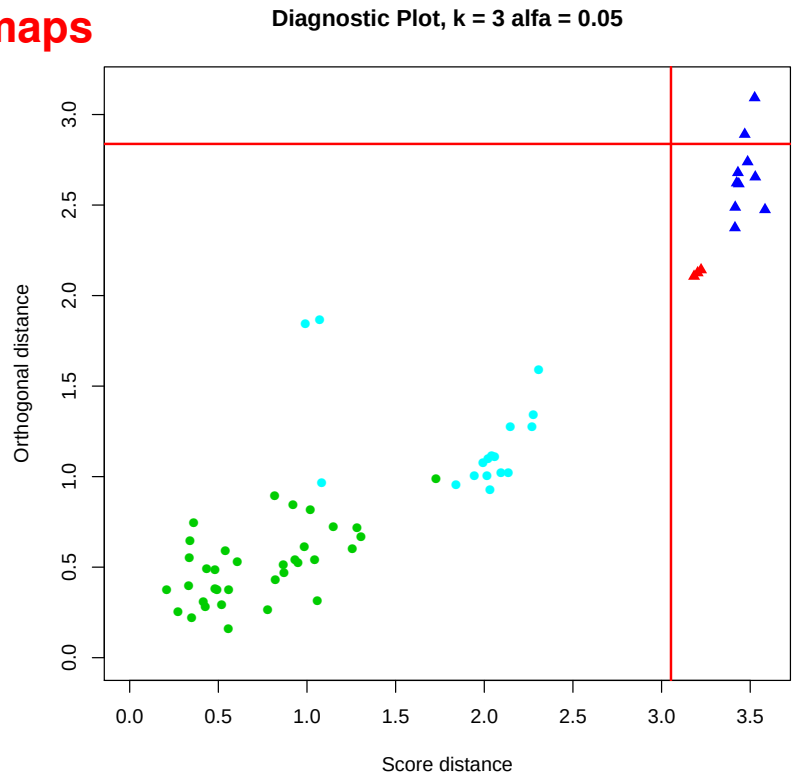
- **BitTorrent mascara as anomalias!**

USA university

Observações ($n = 65$):

- Regulares: 35 **http** + 17 **streaming**
- Contaminação: 10 **snapshots** + 3 **nmaps**
- $k = 3$

	PCAGRID
False Positive	0
Recall	1
Precision	1



Era uma vez...


UM MÉDICO, UM ENGENHEIRO E UM ESTATÍSTICO JUNTARAM-SE PARA TOMAR UM CAFÉ E CONVERSAR...
DIZ O ENGENHEIRO:
PROPUS UM NOVO MÉTODO PARA A DETECÇÃO DE ANOMALIAS NO TRÁFEGO DA INTERNET! AGORA ANDO ÀS VOLTAS PARA TENTAR MOSTRAR AS SUAS POTENCIALIDADES...
PARA O TRÁFEGO REAL, NÃO É FÁCIL ESTABELECER SE UM FLUXO É OU NÃO EFECTIVAMENTE ANÓMALO!
SE CONHECESSE O GROUND TRUTH CONSEGUIRIA CALCULAR MEDIDAS COMO A ENTROPIA E O RECALL E, COM ISSO, MOSTRAR QUE O MEU É UM BOM MÉTODO... ENFIM...
JÁ TENHO UMAS IDEIAS PARA RESOLVER O PROBLEMA...

Era uma vez...


QUE ENGRAÇADO...
A MINHA EQUIPA PROPÔS UMA NOVA TÉCNICA DE DIAGNÓSTICO PARA A DOENÇA QUE ESTAMOS A TRATAR!
VAMOS USAR O MODELO DE CLASSES LACTENTES, JÁ QUE NÃO DISPOMOS DE UM GOLD STANDARD QUE NOS SIRVA DE REFERÊNCIA PARA ESTABELECER A SENSIBILIDADE E ESPECIFICIDADE DO NOSSO MÉTODO!
... O ESTATÍSTICO FICOU CALADO E FOI PARA CASA PENSAR...
AFINAL, O PROBLEMA DO ENGENHEIRO É IGUAL AO PROBLEMA DO MÉDICO!!!
DE FACTO, AS DIFERENÇAS DE LINGUAGEM SÃO BARREIRAS QUE UM ESTATÍSTICO PODE AJUDAR A ULTRAPASSAR!!!

O Problema do Engenheiro...

1. **Identificação de aplicações**:
   - Existem ferramentas de inspecção de pacotes (***Deep Package Inspection*** - **DPI**)
   - Não conseguem identificar todas as aplicações, e.g. **tráfego cifrado**
   - Existe sempre uma percentagem de tráfego que **não** é **identificado** (**Orange DPI**: entre 14% a 39% tráfego não identificado, Pietrzyk et al., 2009)
   - Sabe-se que não são 100% fiáveis. Segundo Pietrzyk et al. (2009), Orange DPI é melhor que Tstat, mas acreditam que **cometem erros**...

O Problema do Engenheiro...

2. **Detecção** automática **de anomalias** em redes de computadores:
   - Invasões a computadores, volumes anormais de tráfego
   - Também tem aplicações em tráfego **rodoviário**
   - **Problema:** Avaliar o **desempenho** de novos métodos de identificação de aplicações ou detecção de anomalias

O Problema do Médico...

- **Testes de diagnóstico** são de grande importância em Medicina visando **classificar** um indivíduo em Doente ou Não Doente

- **Problema:** Avaliar o **desempenho** de novos testes de diagnóstico, na ausência de um teste de referência perfeito (***Gold Standard***)

Indicadores do desempenho de um teste de diagnóstico

- Seja X_i ($i = 1, \ldots, p$) o resultado do i-ésimo teste de diagnostico

$$X_i = \begin{cases} 1, & \text{se o teste dá } \textbf{indicação} \text{ que o indivíduo está doente i.e. } (+) \\ 0, & \text{caso contrário } (-) \end{cases}$$

- Seja Y o verdadeiro estado do indivíduo - Variável Latente

$$Y = \begin{cases} 1, & \text{se o indivíduo está doente} \\ 0, & \text{caso contrário} \end{cases}$$

Indicadores do desempenho de um teste de diagnóstico

As medidas usuais de avaliação do desempenho de um teste de diagnóstico são:

- **Sensibilidade** (Se): probabilidade de um indivíduo doente, D, ser correctamente identificado como doente (**Recall:** probabilidade de uma anomalia, ser correctamente identificada como anomalia), i.e.

$$Se = P(+|D) = P(X = 1|Y = 1)$$

- **Especificidade** (Sp): probabilidade de um indivíduo não doente ($\bar{D}$) ser correctamente identificado como não doente, ($-$), i.e.

$$Sp = P(-|\bar{D}) = P(X = 0|Y = 0)$$

Indicadores do desempenho de um teste de diagnóstico

- **Valor Predictivo Positivo** (*VPP*): probabilidade de um indivíduo diagnósticado como doente, $+$, estar de facto doente (**precision:** probabilidade de um fluxo ser classificado como anómalo, quando de facto é uma anomalia), i.e.

$$VPP = P(D|+) = P(Y = 1|X = 1)$$

Abordagem clássica

Avaliação do desempenho de testes de diagnóstico:

- **Abordagem clássica – Médicos**: comparação do novo teste com um teste de referência idealmente perfeito (***Gold Standard:*** $Se = Sp = 1$)

- Prova-se que o uso de um *Gold Standard* **Imperfeito** como referência conduz a **estimativas enviesadas!**

Abordagem clássica - o que os Médicos sabem...

Como ***Gold Standard imperfeito*** escolhe-se frequentemente o melhor teste disponível

E.g. em parasitologia ainda se escolhe, com frequência, um teste parasitológico, Se<1 e $Sp(X_1) = 1$, como referência

- "*If culture for pertussis is assumed to be <100% sensitive and 100% specific and culture* $(\tilde{Y})$ *is used as the gold standard for assessing the index test* (X)*, then the index test's* ***sensitivity*** *estimate will be* ***unbiased*** *but the* ***specificity*** *estimate will be* ***biased*** *. . .*"

(Baughman et al., 2008)

Verdade se $(\tilde{Y} \amalg X | Y = j)$, onde Y é o *Gold Standard*

Abordagem clássica - o que os Médicos sabem...

Em geral, se $See(X) = P(X = 1|\tilde{Y} = 1)$, $(\tilde{Y} \amalg X|Y = 1)$ e $\eta = P(Y = 1)$ então

$$See(X) = \frac{Se(X)Se(\tilde{Y})\eta + (1 - Sp(X))(1 - Sp(\tilde{Y}))(1 - \eta)}{Se(\tilde{Y})\eta + (1 - Sp(\tilde{Y}))(1 - \eta)}$$

Logo, $Sp(\tilde{Y}) = 1$ implica $See(X) = Se(X)$

De um modo geral,

$$See(X) - Se(X) = \frac{(1 - Sp(\tilde{Y}))(1 - \eta)}{P(\tilde{Y} = 1)}(1 - Se(X) - Sp(X))$$

Logo, $See(X) - Se(X) < 0$ sse $Se(X) + Sp(X) > 1$

Abordagem clássica - o que os Médicos sabem...

"*. . .the* ***specificity*** *estimate will be* ***biased in the direction of lower estimates***" (Baughman et al., 2008)

Verdade se $Se(X) + Sp(X) \geq 1$, $(\tilde{Y} \amalg X | Y = 0)$ e $\eta = P(Y = 0)$

De um modo geral,

$$Spp(X) - Sp(X) = \frac{(1 - Se(\tilde{Y}))\eta}{P(\tilde{Y} = 0)} (1 - Se(X) - Sp(X))$$

Logo, $Spp(X) - Sp(X) \leq 0$ sse $Se(X) + Sp(X) > 1$

Abordagem clássica - o que os Engenheiros sabem...

- ***Ground truth*** perfeito muito difícil de estabelecer
  - Exige o conhecimento da **lista** completa de **anomalias** em conjuntos volumosos de dados. Por sua vez, estes podem ter pouca qualidade

    P. ex. apenas se observam sub-conjuntos de dados (amostragem)

- Muitos dos estudos efectuados em detecção de anomalias e identificação de aplicações carecem de estudos de análise de sensibilidade e de **determinação efectiva** das suas **propriedades** (Ringberg et al., 2008)

Análise de discrepâncias: O que alguns Médicos fazem...

- **Objectivo:** Se o teste de referência é tal que $Sp(\tilde{Y}) = 1$ (e.g. teste parasitológico), sabe-se que se $\{\tilde{Y} = 1\}$ então o indivíduo está parasitado. No entanto, se $\{\tilde{Y} = 0\}$ este pode ser ou não um resultado correcto
- A ideia é "**confirmar**" se os casos $\{\tilde{Y} = 0\}$ e $\{X = 1\}$ são de facto **falsos positivos**
- Aplica-se um novo teste (idealmente, 100% fiável) a este indivíduos. Se o novo teste der positivo considera-se que o resultado $\{\tilde{Y} = 0\}$ estava errado e esta observação passa a ser considerada um **positivo**: $\{\tilde{Y} = 1\}$ e $\{X = 1\}$
- **Actualizam-se** as estimativas de $Se(X)$ e $Sp(X)$

Análise de discrepâncias: o que os Médicos sabem...

- "*. . .discrepant analysis (DA) merely* ***substitutes*** *incorporation bias for imperfect gold standard bias. . .* "
  (Baughman et al., 2008)

- "*DA-based estimate of specificity is typically less biased than that based on culture and that the DA-based estimate of specificity shows little appreciable bias...*" *(Green et al., 1998)*

  "I show that those conclusions are incorrect.... I demonstrate that the concept of discrepant analysis is ***profoundly flawed*** *and* ***unscientific***" (Hadgu, 1999)

Análise de discrepâncias: o que os Médicos sabem...

- De facto, pode provar-se que se $Sp(\tilde{Y}) = 1$ e $(\tilde{Y} \amalg X | Y = j)$ então:

$$Se^{DA}(X) = \frac{Se(X)}{Se(X) + Se(\tilde{Y})(1 - Se(X))}$$

$$Bias = Se^{DA}(X) - Se(X) \geq 0$$

$$Sp^{DA}(X) = \frac{Sp(X)Sp(\tilde{Y})(1-\eta) + (1 - Se(X))(1 - Se(\tilde{Y}))\eta}{(1 - Se(X))(1 - Se(\tilde{Y}))\eta + Sp(\tilde{Y})(1-\eta)}$$

$$Bias = Sp^{DA}(X) - Sp(X) \geq 0$$

Análise de discrepâncias: o que os Médicos sabem...

Alternativas:

- Observar todos os testes de diagnóstico e aplicar MCL (Baughman et al., 2008)
- Admitir mecanismos de censura apropriados e estimar $Se(X)$ e $Sp(X)$ baseado no algoritmo SEM ou usar metodologias bayesianas (Achar et al., 2005; CEB, 2009)

Abordagem clássica - o que os Engenheiros sugerem...

- **Detecção manual de anomalias** *versus* ***deep packed inspection***
  - Ambos estabelecem *ground truth* imperfeitos ou parciais
- **Emulação** - difícil de executar de modo realista e interessante do ponto de vista prático

Análise de discrepâncias: O que alguns Engenheiros fazem...

- Se o procedimento de referência identifica o tráfego como sendo anómalo então é verificado **manualmente** se a anomalia se confirma ou não
- Assim, **actualizam** o *ground truth* imperfeito, que servirá de referência para avaliar o classificador em estudo

Os Engenheiros e o modelo de classes latentes

- **Parâmetros do modelo:** $P(Y = j)$ e $P(X_i = x_i | Y = j)$, $i = 1, \ldots, p$, $j = 0, \ldots, k-1$
- Importando as ideias bem estabelecidas em Medicina, pretendemos aplicar este método para **avaliar métodos** de detecções de anomalias ou classificadores
- **Variáveis observáveis**, X_i, são o resultado da classificação de cada objecto em anomalia ou não anomalia
- **Classes latentes** são a verdadeira classe de tráfego: anomalia ou não anomalia

Comentário Final