# Caracterização e re-identificação de papéis em Redes de Conexão

**Larissa Pinheiro Spinelli**[1], **Daniel R. Figueiredo**[1]

[1]Programa de Engenharia de Sistemas e Computação
Universidade Federal do Rio de Janeiro (UFRJ)
Rio de Janeiro – RJ – Brasil

{larissa, daniel}@land.ufrj.br

***Abstract.*** *Connection Networks are an abstraction to model the exchange of information between entities. In this abstraction, entities are represented by vertices and the exchange of information between two entities is represented by edges. Entities in Connection Networks can have distinct roles which can be related to their functionality. For example, in the Internet Connection Network, entities represented by IP addresses can play the role of client or server. However, many Connection Networks are anonymized in order to omit information concerning the identity and the role of the entities. This paper presents a study of the structural characteristics of the Internet Connection Network as well as a characterization of the different roles played by vertices. Using this characterization, this work proposes techniques to re-identify the role of vertices in anonymized Internet Connection Networks. These techniques use only the structural properties of the network. Finally, the proposed techniques are evaluated and compared to assess their efficiency in re-identifying roles. Numerical results are very promising and indicate that it is possible to re-identify roles with a success rate of over 96%.*

***Resumo.*** *Redes de Conexão são uma abstração para modelar a troca de informação entre um conjunto de entidades. Nesta abstração, entidades são representadas por vértices e a troca de informação entre duas entidadas são representadas por arestas. Entidades em uma Rede de Conexão podem possuir papéis distintos, podendo este estar relacionado com a função desempenhada pela entidade. Por exemplo, na Rede de Conexão da Internet, entidades representadas por endereços IPs podem desempenhar o papel de cliente ou servidor. Entretanto, muitas Redes de Conexão são anonimizadas de forma a omitir informações relacionadas a identidade e o papel das entidades. Este trabalho apresenta um estudo das características estruturais da Rede de Conexão da Internet, assim como a caracterização dos diferentes papéis existentes. Com base nesta caracterização, este trabalho propões técnicas para re-identificação de papéis em Redes de Conexão da Internet que foram anonimizadas. Estas técnicas utilizam-se apenas das propriedades estruturais da Rede de Conexão anônima. Por fim, as técnicas propostas são avaliadas e comparadas entre si para medir a eficiência da re-identificação dos papéis. Resultados numéricos são muito promissores e indicam ser possível re-identificar papéis com taxas de acerto superior a 96%.*

## 1. Introdução

A área de estudo conhecida por Redes Complexas vem possibilitando – pelo desenvolvimento de diversas técnicas e modelos – o entendimento de características e funcionalidades de uma variedade de sistemas em redes presentes na natureza e na sociedade [Albert and Barabási 2002].

Nosso trabalho detém-se ao estudo de um sistema em particular, ao qual denominamos Redes de Conexão. As Redes de Conexão representam troca de informação entre entidades. Um vértice neste sistema representa uma entidade participante da comunicação, e, ao trocar informações, os vértices estabelecem uma aresta entre eles. Um exemplo de Rede de Conexão é a Internet onde os IPs são os vértices da rede e a troca de pacotes entre IPs é representado por uma aresta. Os vértices das Redes de Conexão possuem atributos quanto ao papel desempenhado na troca de informação, ou seja, os vértices são diferenciáveis em tipos. No exemplo da Internet temos que os vértices podem ser denominados cliente se iniciam conexões ou servidor se apenas receberam solicitações de conexões.

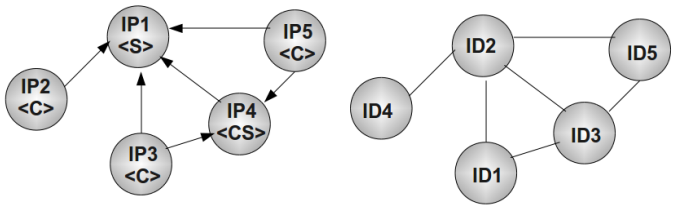
Alguns sistemas possuem informações sensíveis, de cunho pessoal ou sigiloso, recorrendo a anonimização para sua disponibilização pública. A anonimização objetiva a impossibilidade – devido à remoção ou substituição de informações de identidade – de relação, direta ou indireta, entre a instância anônima e a informação real por ela representada. A anonimização das Redes de Conexão leva, por exemplo, a perda de informações quanto ao papel desempenhado pelos vértices e a identidade real destes. Entretanto, muitos estudos recentes mostram a possibilidade da quebra do anonimato ou inferência de informação de identidade, pela exploração das propriedades topológicas das redes.

Este trabalho apresenta uma caracterização das propriedades topológicas das Redes de Conexões, tanto indiscriminadamente, quanto pela distinção de papéis. Inspirando-se nesta caracterização são propostas técnicas que possibilitam a inferência de informações relevantes – como a re-identififcação de papéis, baseando-se apenas na estrutura da rede. Neste trabalho fazemos ainda uma avaliação das técnicas de re-identificação propostas e mostramos que é possível promover a re-identificação de papéis em Redes de Conexões Anônimas com taxas de acerto superior a 96%.

Deste modo, este trabalho está organizado com a seguinte estrutura. Na Seção 2 são definidas, formalmente, as Redes de Conexão e os possíveis papéis nelas identificáveis. Na Seção 3 são apresentados alguns trabalhos relacionados. Na Seção 4 é realizada a caracterização das propriedades topológicas da Redes de Conexões. Na Seção 5 são apresentadas técnicas para a identificação de papéis em Redes de Conexão anonimizadas e, complementarmente, na Seção 6 são apresentados e avaliados resultados da aplicação destas técnicas. Por último, na Seção 7 são feitas as considerações finais sobre o trabalho.

## 2. Rede de Conexão

O conceito de Rede de Conexões foi introduzido no trabalho [Iliofotou et al. 2007] com o nome de Grafo de Dispersão de Tráfego (TDG - *Traffic Dispersion Graph*). TDGs são representações gráficas de várias interações direcionadas (”quem se comunica com quem”) de um grupo de entidades. Nesse contexto a Internet pode ser uma Rede de




(a) Rede de Conexão (b) Rede de Conexão Anônima

**Figura 1. Rede de Conexão e sua Rede de Conexão Anonimizada equivalente**

Conexão no qual os vértices representam entidades com endereços IP distintos e as arestas representam a troca de pacotes entre entidades.

Na construção de uma Rede de Conexão podemos utilizar diversos tráfegos. No nosso estudo consideramos apenas fluxos de comunicação TCP cuja origem ou destino fosse a porta 80. Os vértices deste grafo foram classificados – durante o estudo de caracterização – quanto a inicialização do estabelecimento da conexão TCP entre as entidades. A conexão TCP é assimétrica e, deste modo, entidades que apenas iniciaram conexões foram denominadas como Cliente, as que apenas receberam conexões foram denominadas como Servidor e as que inicializaram e receberam conexões como Cliente/Servidor. Neste trabalho utilizaremos a seguinte notação para referenciar os papéis: Cliente é denotado por C, Servidor por S e Cliente/Servidor por CS. As arestas possuem notação similar denotando os tipos de vértices por elas conectados, por exemplo, arestas que conectam vértices C e S são denotadas por C-S, as que conectam vértices CS e CS serão CS-CS. Na Figura 1(a) vemos um exemplo destas denonimações o IP3 apenas inicia conexões sendo C, o IP1 apenas recebe conexões sendo S e o IP4 tanto inicia como recebe conexões sendo CS. A Rede de Conexão utilizada neste trabalho utiliza apenas o endereço de origem e destino proveniente de pacotes SYN do protocolo TCP para estruturação da rede.

## 2.1. Rede de Conexão Anônima

A anonimização de uma rede pode ser feita utilizando diversas técnicas. Uma delas utiliza-se da desassociação da informação sensível através de um mapeamento um-para-um com identificadores sintéticos [Hay et al. 2008]. A rede anônima gerada é isomorfa à rede identificada que a gerou. Deste modo, um identificador da rede real é sempre mapeado para o mesmo identificador sintético da rede anonimizada bem como, uma aresta existente entre duas instâncias da rede real existirá também entre os vértices correspondentes anônimos. Uma Rede de Conexão Anônima nada mais é uma Rede de Conexão que passou para uma anonimização de identificadores e teve o atributo de papel omitido. A Figura 1 exemplifica este processo. Um problema crítico existente em redes anônimas é a quebra da segurança das informações. Neste sentido, diversos estudos recentes vem mostrando que é possível quebrar o anonimato ou inferir informações de identidade a partir, apenas, da exploração da estrutura da rede anônima. Em Redes de Conexões Anônimas esse problema relaciona-se, por exemplo, a re-identificão de papéis.

## 3. Trabalhos Relacionados

Em [Hay and Srivastava 2006] traces de tráfego real foram utilizados para estruturar Redes de Conexão – definido pelos autores como Grafo de Dispersão de Tráfego (TDGs - *Traffic Dispersion Graphs*). Neste trabalho a análise e utilização de características topológicas é feita com o propósito de classificação de tráfego, e a Rede de Conexões em estudo possui direção quanto ao envio de pacotes. O acréscimo da informação de direção das arestas torna o problema de classificação de papéis proposta neste artigo, trivial.

Em [Narayanan and Shmatikov 2009] é proposto um algoritmo genérico para a re-identificação de vértices em Redes Sociais reais baseado apenas na estrutura da Rede. O algoritmo explora as informações contidas nas arestas como grau e previsão de arestas (*link prediction*) – e não apenas nos dados repassados de cada vértice – para mapear sobreposições entre uma Rede Social anônima alvo e uma Rede Social auxiliar conhecida.

O trabalho desenvolvido em [Pang et al. 2006] apresenta técnicas para inferência da topologia e identificação de servidores em Redes Anonimizadas. [Mahadevan et al. ] mostra um conjunto de característica da topologia AS da Internet.

Em [Meiss et al. 2005] é apresentando um estudo em larga escala do tráfego Web baseado no fluxo de dados de redes. [Kitsak et al. 2010] busca a identificação de um nó que melhor promova a propagação de informação dentro de uma Rede Social.

## 4. Caracterização de Redes de Conexão

Para caracterizar uma Rede de Conexão utilizamos traces públicos de tráfego real do backbone da Internet da base "*The CAIDA Anonymized 2009 Internet Traces*" [Colby Walsworth 2009]. Esta base é composta por traces anonimizados de tráfego passivo – divididos em pedaços correspondentes a 1 minuto de observação - coletada pelos monitores da CAIDA em 2009. Desta base foram utilizados os arquivos "passive-2009/equinix-chicago/20090331" equivalentes a 1 hora de observação consecutiva.

### 4.1. Propriedades

Métricas topológicas são amplamente difundidas para a descrição e comparação de Redes [Albert and Barabási 2002]. Neste trabalho, apenas algumas das métricas mais utilizadas foram avaliadas.

**Grau Médio:** é definido pelas duas mais básicas propriedades de um grafo, o número de vértices e o número de arestas, sendo calculado pela razão entre duas vezes o número de arestas e o número de vértices. O grau médio pode servir como indicativo da conectividade do grafo – grafos com alto grau médio tendem a ser mais conectados e robustos – porém, é tido como um indicativo limitado visto que grafos com diferentes propriedades topológicas podem ter o mesmo grau médio [Albert and Barabási 2002]. A Rede de Conexão em estudo possui $1520327$ endereços de IPs e $2674054$ arestas tendo consequentemente um grau médio de aproximadamente $3,52$.

**Distribuição Empírica de Grau:** é fração de vértices de grau $k$ dada por: $P(k) = n(k)/n$, onde $n(k)$ é o número de vértices com grau $k$ e $n$ o total de vértices. Na Rede de Conexão estudada fica evidente uma relação desigual quanto à distribuição de grau nos vértices. O grau médio do grafo é $3,52$ e o maior grau encontrado é $132900$, ou seja, o maior grau é mais de 37 mil vezes maior que o grau médio. A função de

distribuição cumulativa complementar (*Complementary Cumulative Distribution Function* – CCDF) empírica do grau dos vértices, Figura 4 curva $< todos >$, foi traçada e aproximada a uma distribuição de lei de potência através da utilização do método estatístico de máxima verossimilhança (MLE – *Maximum Likelihood Estimation*) – usado para ajustar dados a um modelo estatístico. Uma lei de potência representa uma relação matemática onde a frequência ou quantidade de um objeto varia de acordo com uma potência de algum atributo. Distribuições de probabilidade que seguem lei de potência representam eventos que, em geral, possuem altas probabilidades para amostras do início da distribuição, como também, amostras muito acima da média da distribuição com probabilidade não desprezível (variando de acordo com uma potência). Redes cujas caudas da distribuição seguem uma lei de potência – ou seja – possuem a forma $P(x) \sim x^{-\gamma}$ são conhecidas como livres de escala [Albert and Barabási 2002]. A CCDF empírica foi aproximada de uma lei de potência com expoente de $\gamma = 2,19$ e erro $\sigma = 4,6 * 10^{-3}$ e, deste modo, comprovamos a grande desigualdade existente entre os graus.

**Distribuição Conjunta Empírica de Grau:** seja $m(k_1, k_2)$ o total de arestas que conectam nós de grau $k_1$ e $k_2$. A distribuição conjunta empírica de grau é a fração de arestas que sejam incidentes sobre vértices de grau $k_1$ e $k_2$:

$$P(k_1, k_2) = \mu(k_1, k_2) * m(k_1, k_2)/2m \text{ , onde } m \text{ é o total de arestas e}$$

$$\mu(k_1, k_2) = \begin{cases} 1 & \text{se } k_1 = k_2 \\ 2 & \text{c.c.} \end{cases}$$

Com a distribuição conjunta de probabilidade é possível estimar informações quanto à vizinhança de um vértice [Albert and Barabási 2002]. Ao observar a Figura 2 vemos

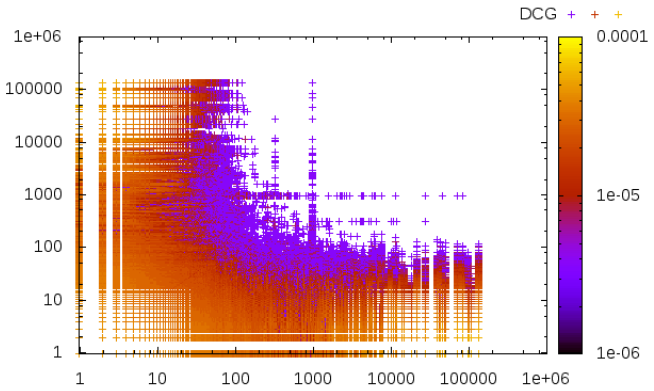



**Figura 2. Histograma da Distribuição Conjunta Empírica de Grau** (Intervalo logarítmo base 2)

que vértices de grau alto são, com grande probabilidade, adjacentes a vértices de grau baixo. Por outro lado, vértices de grau baixo apresentam probabilidades semelhantes de adjacência com graus altos e baixos.

**Componente Conexa:** é definida na teoria do grafos como um sub-grafo conexo maximal. Diz-se que o grafo é conexo quando ele possui apenas um componente conexo

composto pelo grafo inteiro. A Rede de Conexão em estudo não é conexa sendo, entretanto, dominada por uma Componente Conexa Gigante (GCC - *Giant Connected Component*) que possui $1499870$ vértices, o equivalente a 98,65% do total de vértices restando apenas 1,35% nas demais componentes. A Rede possui um total de $7869$ componentes conexos nos quais 78,88% tem tamanho $2$ (menor tamanho possível), 13,09% tamanho 3, 4,43% tamanho $4$ e, consequentemente, menos de 3,6% tem tamanho maior que $4$. Esta relação desigual entre o tamanho da componente conexa e a quantidade de componentes está expressa na Figura 3 onde a CCDF empírica do tamanho da componente conexa evidencia tal diferença, de onde foi calculado o expoente $\gamma = 2,22$ com o qual temos um erro $\sigma = 1,3 * 10^{-2}$. **Distância:** a distância entre dois vértices é definida como o menor

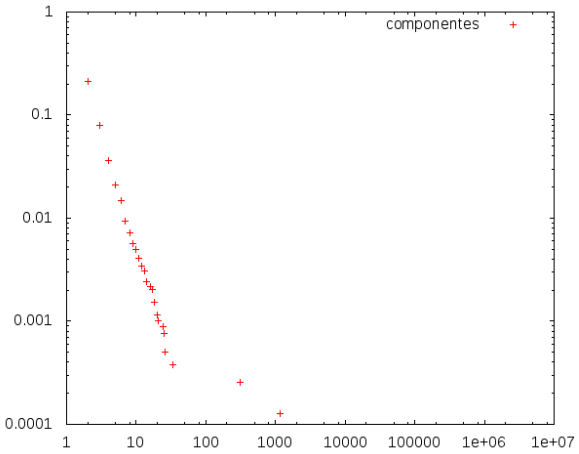



**Figura 3. Distribuição do tamanho das componentes**

caminho entre eles. A distribuição empírica da distância $d(x)$ representa a fração de pares de vértice que estão a uma distância de $x$ saltos do outro. A maior distância do grafo representa o diâmetro deste.

**Clusterização:** o coeficiente de clusterização representa a porcentagem de triângulos sobre o total de triplas conectadas em todo o grafo [Albert and Barabási 2002]. Esta métrica tem implicações práticas pois expressa a robustez do grafo. Muitas das redes complexas reais estudadas possuem alto coeficiente de clusterização, porém, para a rede em estudo – surpreendentemente – a clusterização é nula. Vale ressaltar que a clusterização nula significa que não há ciclo de tamanho três na Rede de Conexões avaliada. Entretanto, esta rede possui ciclos de outros tamanhos.

## 4.2. Caracterização por Papéis

Nesta seção faremos a caracterização da Rede de Conexões quanto aos papéis desempenhados pelos vértices. A observação das propriedades topológicas do Grafo quanto a estes papéis é essencial para a caracterização e diferenciação destes. A partir desta análise é possível, então, propor técnicas para a classificação de papéis em Redes de Conexão anonimizadas.

**Grau Médio:** é calculado pelas soma do grau dos vértices de um dado papel dividido pelo total de vértices deste tipo. A Rede em estudo possui $1399690$ vértices clientes – 92,07% do total – que possuem o grau médio de aproximadamente $1,89$. Para o tipo servidores foram classificados $116509$ vértices (7,66%) e o grau médio é aproximadamente

$21,86$. Já para o tipo cliente/servidor o grafo possui apenas $4128$ nós (0,27%) e o grau médio de $37,90$.

**Distribuição Empírica de Grau:** seja $n_1(k)$ o número de vértices com grau k do $tipo_1$. A distribuição empírica de grau dos vértices por tipo corresponde à fração de vértices de um determinado tipo com grau $k$: $P(k|t_1) = n_1(k)/n_1$ . A Figura 4 mostra a CCDF empírica de grau para os diferentes papéis na Rede. Para o tipo C foi estimado um expoente $\gamma_c = 2,48$ com o qual temos um erro $\sigma_c = 5,5 * 10^{-3}$, para o tipo S $\gamma_s = 1,73$ e $\sigma_s = 7,4 * 10^{-3}$ e para o tipo CS $\gamma_{cs} = 2,05$ com o qual temos um erro $\sigma_{cs} = 3 * 10^{-2}$. Deste modo vemos que não só a distribuição empírica de grau da rede segue uma lei de potência, como também, cada distribuição empírica de grau por tipo de vértice estudada segue lei de potência. É interessante observar que as distribuições de grau para S e CS tem um expoente menor que o expoente da distribuição C e, em consequência, possuem uma distribuição de grau ainda mais desigual. Na figura 4 fica

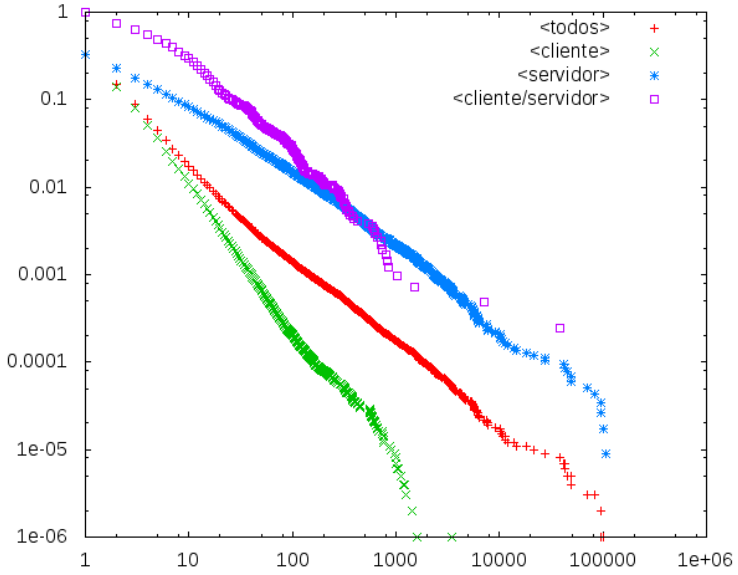
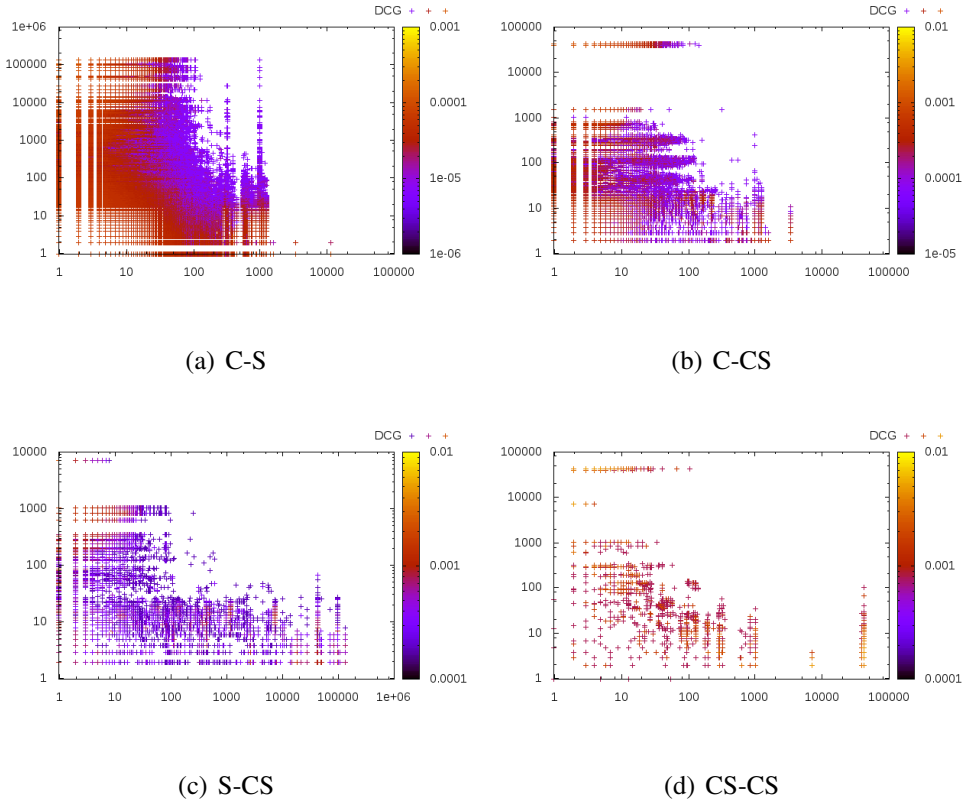



**Figura 4. Distribuição Empírica de Grau**

evidente a diferença entre as caudas das distribuições. O maior grau de um vértice cliente é 11411. Para este valor temos ainda que 0,19% dos vértices do tipo servidor, ou seja, mais de 220 vértices, possuem grau maior que este. Para os vértices do tipo cliente/servidor esse valor chega a 0,9%, porém representa uma quantidade de apenas 37 vértices com grau superior. O maior grau de um vértice cliente/servidor é de 42384, para o qual ainda temos aproximadamente 0,01% de vértices do tipo servidor com grau superior, cerca de 10 vértices. O maior grau da rede é do tipo servidor, com grau de 132900 que é 3 vezes maior que o maior grau do tipo cliente/servidor e 11 vezes maior que o maior grau do tipo cliente.

**Arestas:** A Rede de Conexões em estudo, devido ao modo como é composta, possibilita que existam apenas arestas entre determinados tipos de vértices. As arestas possíveis são: C-S, C-CS, S-CS e CS-CS. Destas observou-se as seguintes quantidades dentro na rede:

- C-S: 2520009 arestas (94,24%)




(a) C-S (b) C-CS

(c) S-CS (d) CS-CS

**Figura 5. Histograma da Distribuição Conjunta Empírica de Grau por Papéis** (Intervalo logarítmo base 2)

- CS-C: 124717 arestas (4,66%)
- CS-S: 26932 arestas (1,01%)
- CS-CS: 2396 arestas (0,09%)

As proporções de arestas nos revelam que a adjacência mais comum na rede é a de Clientes com Servidores. Outra maneira de interpretar esses dados diz respeito à frequência relativa da classificação dos vértices adjacentes a um determinado tipo. Neste sentido, os vértices do tipo S tem 98,94% de vizinhos do tipo C e apenas 1,06% do tipo CS. Já os vértices do tipo C tem 95,28% de vizinhos do tipo S e 4,72% do tipo CS. Vértices CS tem 80,96% de vizinhos C, 17,48% S e apenas 1,56% CS.

**Distribuição Conjunta Empírica de Grau por papéis:** representa a fração relativa de arestas que conectam um vértice de grau $k_1$ do $tipo_1$ e de grau $k_2$ do $tipo_2$. Como arestas do tipo C-C e S-S não são possíveis, a distribuição conjunta para estas arestas é nula. As demais distribuições empíricas conjuntas estão representadas na Figura 5. A Figura 5 nos revela algumas tendência de vizinhança da rede. Em 5(a) vemos que S de grau alto estão conectados na sua maioria com C de grau baixo e C de grau alto conectam-se, em geral, a S de grau baixo. Tal padrão – relação entre grau alto e baixo – é repetido também na Figura 5(b) e menos intensamente nas Figuras 5(c) e 5(d).

## 5. Re-identificação de papéis

A inferência de informações relevantes provenientes apenas da estrutura de Redes anonimizadas tem sido alvo de diversos trabalhos [Hay et al. 2008] e [Pang et al. 2006]. Nesta

seção apresentamos algumas técnicas, propostas neste trabalho, para a re-identificação de papéis (cliente, servidor, cliente/servidor) em Redes de Conexão anonimizadas. As técnicas apresentadas exploram apenas as propriedades estruturais destas redes descritas e avaliadas na seção 4 deste trabalho. Todas as técnicas propostas preocupam-se em gerar classificações consistentes, ou seja, classificações possíveis, observadas as restrições de relação de adjacência. Uma classificação consistente para uma Rede de Conexão não permite que existam arestas do tipo C-C ou S-S.

### 5.1. Classificador de Ordem Decrescente

Esta técnica tenta explorar, de modo simples, a relação entre as caudas da distribuição empírica de grau dos diferentes tipos de vértice. Podemos inferir que os vértices de maior grau da rede serão do tipo servidor, de acordo com a avaliação na seção 4.2. Além disso, esta técnica de classificação respeita as possíveis relações de adjacências entre os vértices, não permitindo que um vértice do tipo cliente ou servidor sejam adjacentes a vértices do mesmo tipo. A idéia geral deste classificador é percorrer os vértices da Rede de Conexão em ordem decrescente quanto ao grau, observar os vizinhos já classificados e atribuir uma classificação consistente. O esquema desta técnica está descrito abaixo:

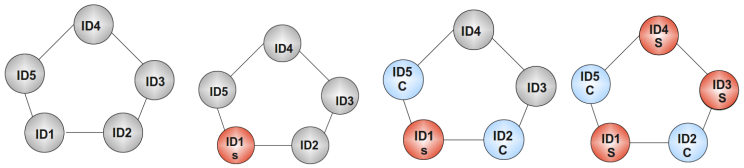
```
Classificador OrdemDecrescente (Maiores:FILA, Grafo: GRAFO)
   INTEIRO: Cli, Serv
   VERTICE: e, v
   Para v em Vertices(Grafo):  rotulo[v] := 0
   Para e em Maiores:
      Cli, Serv := 0
      Para v em Vizinhos(Grafo, e):
         caso rotulo[v] = C: Cli ++
         caso rotulo[v] = S: Serv ++
      Fim Para
      se Serv=0 rotulo[e] :=  S
      se nao, se Cli=0 rotulo[e] := C
      se nao rotular[e] := CS
   Fim Para
Fim Classificador
```

### 5.2. Classificador BFS

A Busca em Largura (BFS – *Breadth-First Search*) é um dos algoritmos mais simples e utilizados para se percorrer um grafo. Este método explora sistematicamente as arestas de um Grafo a partir de um vértice, até descobrir todos os vértices acessíveis por este iniciador [Cormen et al. 2001].

A idéia geral deste classificador parte da combinação da propagação de uma BFS com algumas constatações provenientes da análise das Redes de Conexão, feita na seção 4, como a inferência quanto à classificação do vértice de maior grau como servidor e a relação de adjacências possíveis entre os vértices. As proporções das adjacências existentes justificam o fato da utilização da BFS. Dado que 98,94% dos vértices adjacentes a um vértice servidor são clientes e que 95,28% dos vértices adjacentes a um vértice cliente são servidores, tenta-se classificar vértices como servidores ou clientes alternadamente em cada onda de propagação. Adicionalmente, esta técnica utiliza-se também da constatação da existência de ao menos um vértice cliente/servidor em ciclos de tamanho ímpar.

A existência de ao menos um vértice cliente/servidor em um ciclo ímpar é resultado das relações de adjacência possíveis. De outra maneira, se pudéssemos classificar em




(a) 1 (b) 2 (c) 3 (d) 4

**Figura 6. Detecção de CS em ciclo ímpar.**

um ciclo ímpar vértices apenas com atribuições de cliente ou servidor, chegaríamos a uma inconsistência. Ao começar a classificação com um iniciador dentro do ciclo, devido a alternâcia em cada passo entre as classificações (cliente ou servidor), haverá um ponto no qual as classificações se interceptam e, neste ponto, havéra dois vértices classificados com o mesmo tipo C-C ou S-S, chegando a inconsistente mencionada. A figura 6 exemplifica esta constatação. Esta técnica proposta encontra-se apresentada no quadro a seguir:

```
Classificador BFS (Maiores:FILA, Grafo:GRAFO)
  FILA: Fila := pegarPrimeiros(1, Maiores)
  BOLEANO: aux
  VERTICE: n, v
  Para v em Vertices(Grafo):
      rotulo[v] := 0
      visita[v] := 0
  Fim Para
  Para n em Fila: camada[n] := 0
  Enquanto Fila não vazia:
      n := saiFila(Fila)
      aux := FALSO
      se camada[n] é Par
          rotulo[n] := S
      se  camada[n] é ímpar
          rotulo[n] := C
      Para v em Vizinhos(Grafo, n):
        se visita[v]= 0
           visita[v]:=1;
              entrarFila(Fila, v);
              camada[v] := camada[n]+1
        se visita[v] = 1
              se rotulo[v] = rotulo[n]
                  aux := VERDADEIRO
      Fim Para
      se aux
          rotulo[n] := CS
  Fim enquanto
Fim Classificador
```

### 5.3. Classificador com múltiplas BFS

O Classificador BFS, ao classificar um vértice errado, propaga o erro por toda a árvore que possui o vértice classificado erroneamente como pai. Apesar da grande porcentagem de adjacências de C-S, um pequeno erro de classificação no início da árvore BFS pode gerar grandes propagações de erro dentro da rede. A classificação com múltiplas BFS explora todas as características já abordadas pela classificação BFS (simples). A técnica utiliza os $n$ vértices de maior tamanho como iniciadores tentando classificá-los como servidores. Esta técnica possibilita a diminuição de erros no início da propagação da classificação, a detecção de vértices CS fora de ciclos ímpares e a diminuição do número de rodadas.

O esquema da técnica de múltiplas BFS só difere do esquema de uma única BFS pela inicialização da Fila como pode ser visto a seguir:

```
FILA: Fila := pegarPrimeiros(n, Maiores)
```

## 6. Avaliação dos Classificadores

A avaliação de sistemas de classificação é feita de forma experimental observando a eficácia do classificador, ou seja, sua capacidade de classificar corretamente as instâncias avaliadas. Segundo [Baeza-Yates and Ribeiro-Neto 1999] uma das maneiras de se calcular esta efetividade é utilizando-se contadores extraídos de uma matriz de contingência e utilizar, por exemplo, as medidas clássicas de precisão e abrangência.

A matriz de contingência apresenta contadores relativos às quantidades de objetos classificados como pertencentes ou não a uma determinada classe, pelo classificador especialista. Observando-se uma classe particular $x$ podemos obter 4 contadores da matriz de contingência, que são: TP quantidade de objetos classificados corretamente como $x$, FP quantidade de objetos classificados erroneamente como $x$, TN quantidade de objetos corretamente não classificados como $x$, FN quantidade de objeto erroneamente não classificados como $x$. A partir desses contadores podemos facilmente calcular a precisão e a abrangência. A precisão consiste na probabilidade da classificação estar correta (TP), dado que o objeto havia sido classificado como $x$ (TP+FP), ou seja, a precisão é TP/(TP+FP). A abrangência mede a probabilidade de um objeto, tirado ao acaso, ser classificado como $x$ (TP+FN) e que esta classificação esteja correta (TP), ou seja, a abrangência corresponde a: TP/(TP+FN).

Para simplificar a avaliação, as medidas de precisão e abrangência podem ser combinadas, por exemplo, utilizando a medida F (F-measure) [Baeza-Yates and Ribeiro-Neto 1999]. A medida F pode ser calculada da seguinte forma:

$$F_\alpha = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

onde $\alpha$ é um fator da importância relativa da precisão e da abrangência. Em nossas avaliações utilizaremos $\alpha = 0,5$

Outra medida de avaliação para os classificadores é a complexidade computacional de pior caso [Cormen et al. 2001]. Esta medida pode ser obtida pela inspeção do pseudo-código de cada classificador. Inspecionando o Classificador de Ordem Crescente vemos que este recebe os vértices ordenados, a ordenação tem complexidade $O(n \log n)$, onde $n$ é o número de vértices do grafo. Além disso, este classificador percorre a lista de vértices uma única vez e, para cada vértice analisado, são inspecionados todos os seus vizinhos. Esta quantidade de iterações é dada por: $\sum_{i=1}^{n} g(i)$, onde $n$ é o número de vértices do grafo e $g(i)$ é o grau do vértice $i$. O resultado deste somatório é igual a $2m$, onde $m$ é o número de arestas do grafo [Cormen et al. 2001]. Temos assim que a complexidade deste primeiro classificador é $O(n \log n + m)$. O classificador BFS, por sua vez, requer a identificação do vértice de maior grau, o que é feito com uma busca simples de complexidade $O(n)$. Adicionalmente, a partir deste vértice inicial, percorre-se a lista de vértices

apenas uma vez inspecionando os vizinhos de cada vértice. A complexidade do classificador BFS é portanto $O(n+m)$ [Cormen et al. 2001]. O classificador com múltiplas BFS percorre a lista de vértices uma única vez, pois não há sobreposição nas propagações das múltiplas BFS. Para a identificação dos iniciadores destas BFS, entretanto, é necessário obter a lista dos $k$ maiores vértices. Como $k$ é uma constante correspondente ao número de iniciadores, poderíamos realizar, por exemplo, $k$ buscas simples. Deste modo a complexidade deste classificador é $O(n + m)$. Comparativamente, podemos concluir que a complexidade dos classificadores BFS será sempre melhor ou igual a complexidade do Classificador de Ordem Crescente.

### 6.1. Resultados

Para uma maior compreensão dos resultados, todos os classificadores tiveram seus resultados classificados em métricas de precisão e abrangência, tanto em uma abordagem geral – observando-se todos os tipos de vértices, como também em uma abordagem específica por tipo de vértice. Vale ressaltar que na abordagem indiferente quanto ao tipo (todos) as métricas de precisão, abrangência e Medida-F são equivalentes e representam a taxa simples de acertos.

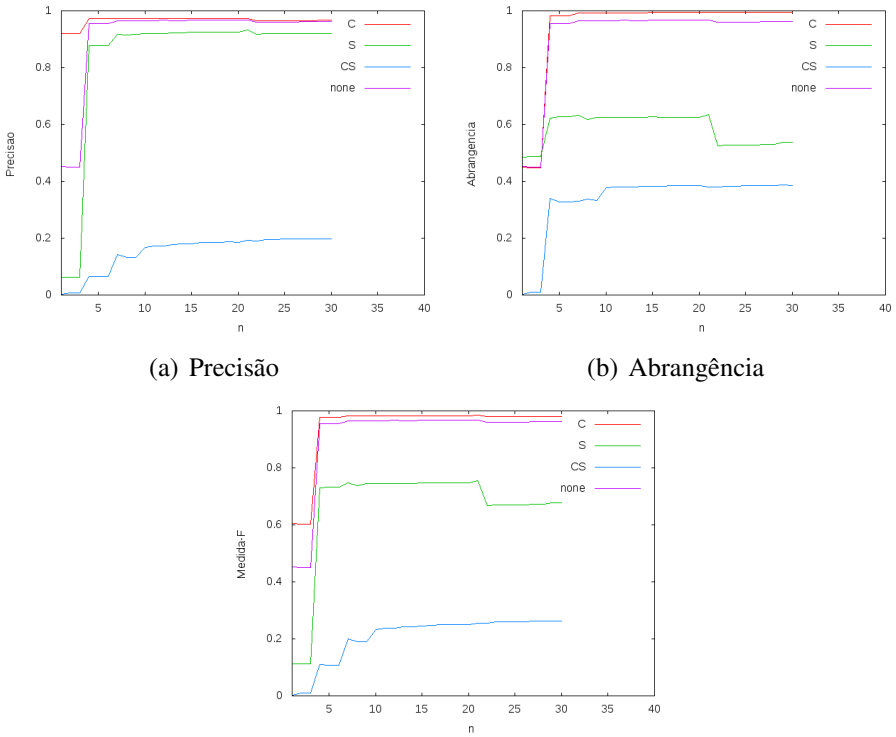
A Tabela 1 mostra os resultados obtidos pelos classificadores. Utilizaremos os resultados de $n = 5, 10, 20, 30$ para o classificador de múltiplas BFS.

| | Precisão – C | Precisão – S | Precisão CS | Abrangência – C | Abrangência – S | Abrangência – CS | Medida-F - todos |
|---|---|---|---|---|---|---|---|
| **Ordem Decrescente** | 0,9671 | 0,3879 | 0,1365 | 0,9293 | 0,5556 | 0,4198 | 0,9016 |
| **BFS** | 0,9194 | 0,0634 | 0,0026 | 0,4510 | 0,4860 | 0,0037 | 0,4523 |
| **5-BFS** | 0,9731 | 0,8787 | 0,0647 | 0,9826 | 0,6270 | 0,0647 | 0,9557 |
| **10-BFS** | 0,9731 | 0,9200 | 0,1669 | 0,9933 | 0,6248 | 0,3763 | 0,9656 |
| **20-BFS** | 0,9735 | 0,9262 | 0,1860 | 0,9941 | 0,6261 | 0,3812 | 0,9678 |
| **30-BFS** | 0,9731 | 0,9210 | 0,1735 | 0,9936 | 0,6243 | 0,3795 | 0,9668 |

**Tabela 1. Avaliação das Técnicas**

As técnicas propostas apresentam, em geral, melhor eficiência para a classificação de clientes, seguido por servidor e, por último, de cliente/servidores. A técnica de BFS, apesar de apresentar uma boa precisão para a classificação de clientes, possui uma eficiência muito baixa. As técnicas de Ordem Descrecente apresentam bons resultados para a classificaçao de clientes e médios resultados para a classificação nos demais tipos, tendo uma avaliação geral boa. Os melhores resultados obtidos foram para o classificador de múltiplas BFS, apresentando excelentes resultados para o tipo cliente, bons resultados para o tipo servidor e resultados médios para o tipo cliente/servidor, tendo portanto uma ótima avaliação final. Esta técnica de classificação só perde para na abragência da avaliação de cliente/servidores para o classificador de Ordem Decrescente.

Entre os classificadores BFS – com única propagação ou múltiplas – é notável o aumento na eficiência da classificação daqueles que utilizam múltiplos iniciadores em relação ao com iniciador único. Os classificadores com múltiplas BFS foram experimentados com diversas variações quanto ao número $n$ de iniciadores. A Figura 7, por exemplo, representa os resultados obtidos com a variação de iniciadores com $n$ entre 1 e 30. Os resultados obtidos observando-se apenas o acréscimo de um ou dois iniciadores poderiam desacreditar a técnica de classificação com múltiplas BSF, pois a taxa de acerto passa de 45,23% para 44,93% para $n = 2$. Porém ao utilizar 4 iniciadores observa-se um




(a) Precisão (b) Abrangência

(c) Medida-F

**Figura 7. Avaliação do Classificador com n múltiplas BFS**

aumento significativo na eficiência da classificação cuja taxa de acertos salta de 44,88% para 95,54% – um aumento de cerca 112,88%. Depois deste salto o acréscimo de mais iniciadores desenha-se como uma aparente função crescente de leves oscilações.

Para entender melhor os resultados obtidos com a utilização de classificadores com múltiplas BFS foram observadas as distâncias existentes entre esses vértices iniciadores. A Tabela 2 representa as distâncias existentes entre o grupo dos 10 vértices de maior grau ordenados de modo decrescente. Adicionalmente apresentamos também a altura da árvore gerada pela BFS de cada um deles. Desta tabela vemos que o quarto maior vértice estava a uma distância ímpar do vértice iniciador no caso de $n = 1$. Tal fato ocasiona – pela atribuição de rótulos da rodada deste algoritmo – a classificação errônea deste vértice, e da árvore subsequente que possui ele como pai. O mesmo acontece também nos vértices de sétimo, oitavo e décimo maior grau. Observe que ao utilizar $n = 4$ garantimos que todos os 10 maiores vértices serão atingidos em rodadas pares e serão classificados como servidores.

## 7. Considerações finais

As contribuições chaves deste trabalho foram a avaliação e caracterização das propriedades topológicas de uma Rede de Conexão retratando o tráfego TCP na porta 80, como também, o desenvolvimento de técnicas para a re-identificação de papéis em Redes de Conexões Anonimizadas baseado apenas em sua estrutura.

A melhor técnica proposta avaliada – 30 múltiplas BFS – possui uma taxa de acerto de 96,78% comprovando que é possível identificar papéis em Redes de Conexão

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Max |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 0 | 2 | 2 | 3 | 2 | 2 | 3 | 3 | 2 | 3 | 13 |
| **2** | 2 | 0 | 2 | 3 | 2 | 2 | 3 | 3 | 2 | 3 | 14 |
| **3** | 2 | 2 | 0 | 3 | 2 | 2 | 3 | 3 | 2 | 3 | 14 |
| **4** | 3 | 3 | 3 | 0 | 3 | 3 | 2 | 2 | 3 | 2 | 12 |
| **5** | 2 | 2 | 2 | 3 | 0 | 2 | 3 | 3 | 2 | 3 | 13 |
| **6** | 2 | 2 | 2 | 3 | 2 | 0 | 3 | 3 | 2 | 3 | 13 |
| **7** | 3 | 3 | 3 | 2 | 3 | 3 | 0 | 2 | 3 | 2 | 12 |
| **8** | 3 | 3 | 3 | 2 | 3 | 3 | 2 | 0 | 3 | 2 | 14 |
| **9** | 2 | 2 | 2 | 3 | 2 | 2 | 3 | 3 | 0 | 3 | 13 |
| **10** | 3 | 3 | 3 | 2 | 3 | 3 | 2 | 2 | 3 | 0 | 12 |

**Tabela 2. Distância entre os 10 vértices de maior grau**

Anônimas.

## Referências


Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(1):47–97.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley.

Colby Walsworth, Emile Aben, k. c. D. A. (2009). The caida anonymized 2009 internet traces - ¡dates used¿.

Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. (2001). *Introduction to Algorithms, Second Edition*. McGraw-Hill Science/Engineering/Math.

Hay, M., Miklau, G., Jensen, D., Towsley, D. F., and Weis, P. (2008). Resisting structural re-identification in anonymized social networks. *VLDB*, 1(1):102–114.

Hay, M. and Srivastava, S. (2006). Privacy and anonymity in graph data.

Iliofotou, M., Pappu, P., Faloutsos, M., Mitzenmacher, M., Singh, S., and Varghese, G. (2007). Network monitoring using traffic dispersion graphs (tdgs). In *Proceedings of the 7th ACM SIGCOMM Internet Measurement Conference*, pages 315–320, New York, NY, USA. ACM.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., and Makse, H. A. (2010). Identifying influential spreaders in complex networks. cite arxiv:1001.5285 Comment: 31 pages, 12 figures.

Mahadevan, P., Krioukov, D., Dimitropoulos, X., Huffaker, B., Fomenkov, M., kc claffy, and Vahdat, A. The internet as-level topology: Three data sources and one definitive metric.

Meiss, M., Menczer, F., and Vespignani, A. (2005). On the lack of typical behavior in the global web traffic network. In *WWW '05: Proceedings of the 14th international conference on World Wide Web*, pages 510–518, New York, NY, USA. ACM.

Narayanan, A. and Shmatikov, V. (2009). De-anonymizing social networks. *Security and Privacy, IEEE Symposium on*, 0:173–187.

Pang, R., Allman, M., Paxson, V., and Lee, J. (2006). The devil and packet trace anonymization. *SIGCOMM Comput. Commun. Rev.*, 36(1):29–38.