

Estimação de Matrizes de Tráfego Origem-Destino Utilizando Algoritmo Genético

Fabiano C. Ribeiro¹, Arnaldo N. da Silva², J. E. Bessa Maia¹, Jorge L. C. Silva¹,
Paulo R. F. Cunha²

¹Departamento de Estatística e Computação - Universidade Estadual do Ceará (UECE)
Av. Paranjana, 1700 - 60740-903 – Fortaleza, CE – Brazil

²Centro de Informática – Universidade Federal de Pernambuco (UFPE)
Caixa Postal 7851 – 50732-970 – Recife, PE – Brazil

{fcr,jmaia,jlcs}@larces.uece.br, {ans2,prfc}@cin.ufpe.br

Abstract. *The accurate estimating of the traffic matrices origin-destination from the traffic measured in links of an IP network is a complex problem for which has not been found suitable solution. This paper applies genetic algorithms (GA) to the problem of estimating and compares the results with those obtained by three other techniques already applied: Linear Programming, Bayesian Inference and approximation for Expectation Maximization (EM) algorithm. Besides the use of GA, another contribution of this work is an initialization algorithm of parameters that decreases the computational effort and aids the convergence for expected result. The experiments show a better performance of the genetic algorithm in terms of the estimated error.*

Resumo. *A estimação precisa da matriz de tráfego origem-destino a partir do tráfego medido nos enlaces de uma rede IP é um problema complexo para o qual ainda não foi encontrada solução satisfatória. Este trabalho aplica algoritmos genéticos (AG) a este problema de estimação e compara os resultados com aqueles obtidos através de três outras técnicas aplicadas anteriormente: Programação Linear, Estimação Bayesiana e aproximação pelo algoritmo Esperança-Maximização (EM). Além da utilização de AG, uma outra contribuição deste trabalho é um algoritmo de inicialização de parâmetros que diminui o esforço computacional e auxilia a convergência para o resultado esperado. Os experimentos apontam para um melhor desempenho do algoritmo genético do ponto de vista do erro de estimação.*

1. Introdução

Matrizes de Tráfego Origem-Destino (MT) refletem o volume de tráfego que flui entre todos os possíveis pares de nós de uma rede. Uma MT tem múltiplas utilizações na engenharia de tráfego e operação de redes, incluindo balanceamento de carga, configuração de protocolo de roteamento, dimensionamento, implementação de políticas de QoS, provisionamento e estratégias para superar falhas. A localização e o tamanho dos fluxos que compõem a MT são informações importantes para o planejamento do crescimento de uma rede e diagnósticos de problemas. [Medina et al 2002]

Os métodos atuais de medição direta dos fluxos que compõem a matriz de tráfego são dispendiosos ou quase sempre inviáveis. Por outro lado, a partir da MIB (Management Information Base) do SNMP (Simple Network Management Protocol), as medições dos fluxos agregados dos enlaces ponto-a-ponto são facilmente obtidas. Assim, a abordagem comumente utilizada para obter a MT é estimá-la a partir das medições de tráfego nos enlaces. Entretanto este processo de estimação não é trivial.

As dificuldades desta estimação surgem de várias origens. Primeiro, as redes em geral apresentam baixa conectividade de forma que o número de enlaces resulta ser menor que o número de fluxos origem-destino (OD). Isto dá origem a um problema inverso sub-determinado que permite infinitas soluções. Segundo, pode não ser fácil obter a matriz de roteamento correspondente aos fluxos medidos.

Existem várias técnicas propostas para estimação de uma MT dentro do contexto da Internet, redes de transporte e redes de telecomunicações. Este trabalho foca o problema da estimação de tráfego em rede IP, baseada nos algoritmos genéticos (AG). Que seja do nosso conhecimento, algoritmo genético não foi aplicado anteriormente a este problema. No entanto, o AG aplicado ainda é incipiente, pois usa somente o operador de mutação, ou seja, é do tipo mais simples: subida de encosta (Hill Climbing). Uma classe de algoritmos de busca estocástica é, em geral, muito dependente de uma boa solução inicial. Outra contribuição original deste trabalho é um algoritmo de inicialização de parâmetros que se mostrou eficiente na redução do tempo computacional e do erro médio de estimação.

As técnicas utilizadas anteriormente para estimação da MT são de duas classes: baseadas em programação linear e baseadas em métodos estatísticos.

Dentre essas classes, o trabalho de Medina et al (2002) faz um estudo comparativo entre a abordagem baseada em Programação Linear (PL) [Goldscmidt 2000], Técnicas de Inferência Bayesianas [Tebaldi e West 1998] e Maximização de Expectativa [Cao et al 2000], onde são propostos, também novos direcionamentos baseado em seus resultados. Tebaldi e West (1998) desenvolvem um framework teórico de variantes do problema de fluxo origem-destino e utiliza modelos Bayesianos para análise de tráfego de rede em problemas de inferência [Vardi 1996] sobre tráfego direcionado entre pares de origens e destinos em redes. Podemos ver ainda que em [Zhang et al 2003(a)] é utilizado teoria da informação como abordagem para esse problema. Soule et al (2004) adotam um estimador da variância das matrizes de tráfego. Em [Soule et al 2005] é feita uma avaliação do equilíbrio entre a adoção de medições, o uso de inferência e modelagem para a estimação de matrizes de tráfego. Teixeira et al (2005) avaliam o efeito de mudanças no roteamento na variação das matrizes de tráfego.

O resultado obtido nesse trabalho foi melhor do que os métodos estatísticos que requerem um grande poder de processamento de informações e também melhor do que a Programação Linear descrita em [Goldscmidt 2000]. Silva et al (2007) propõem um novo conjunto de restrições para o uso em programação linear que limita o conjunto de soluções viáveis para estimar medidas de tráfego entre pares de nós, proporcionando melhores resultados que os de [Medina et al 2002].

Para a validação deste trabalho foi utilizada a topologia de rede apresentada por Tebaldi e West (1998) e Medina et al (2002) e também os dados reais de tráfego da rede GÉANT [Uhlig et al 2006].

Este trabalho está organizado como a seguir. A seção 2 traz uma revisão dos métodos de estimação de MT relevantes ao estudo. A metodologia do trabalho e a descrição das topologias de redes usadas estão presentes na seção 3. A seção 4 descreve os resultados obtidos a partir do estudo e as seções 5 e 6 apresentam respectivamente as conclusões e agradecimentos e por fim a bibliografia.

2. Revisão dos Métodos

2.1. Estimação de Matrizes de Tráfego

O tráfego observado em um enlace da rede é formado pela superposição de fluxos OD, onde cada um desses fluxos consiste de todo o tráfego que entra em um ponto de ingresso comum e sai em outro ponto da rede. Seja c o número de pares OD de uma rede formada por n nós. Então $c = n \times (n - 1)$. Conceitualmente, os fluxos de tráfego podem ser representados em uma matriz X , onde a quantidade de dados transmitida do nó i para o nó j é denotada pelo elemento x_{ij} . Entretanto, é mais conveniente utilizar uma representação de vetor para representar os fluxos, onde os pares OD ficam ordenados na forma x_j , $j = 1 \dots c$. Seja $Y = (y_1, y_2, \dots, y_r)$ o vetor das medidas do tráfego nos enlaces, onde y_l é a medida do tráfego no enlace l , e r representa o número de enlaces da rede. O relacionamento entre os enlaces e os fluxos OD pode ser evidenciado através de uma matriz de roteamento A do tipo $\{0,1\}$, onde as linhas representam os enlaces da rede e as colunas representam os pares OD. A matriz A tem tamanho r (enlaces) por c (fluxos), onde $A_{ij} = 1$ se o fluxo j passa pelo enlace i e toma o valor zero caso contrário.

A relação entre o vetor de pares ordenados X e o vetor de tráfego de enlaces Y é dada por:

$$Y = AX \quad (1)$$

A matriz de roteamento pode ser obtida através de pesos de enlaces utilizando os algoritmos OSPF (Open Shortest Path First) ou IS-IS (Intermediate System) e calculando o caminho mais curto entre todos os pares OD. Os dados dos enlaces estão disponíveis através do SNMP. Assim, o problema principal consiste em estimar X , ou seja, encontrar um conjunto de fluxos OD que poderá reproduzir Y o mais próximo possível do real. Este problema, representado pela equação (1), é indeterminado, pois em praticamente todas as redes o número de pares OD é muito maior que o número de enlaces na rede, $r \ll c$. Isto significa que há um número infinito de soluções possíveis para o vetor X .

2.2. Modelo de Programação Linear (PL)

O problema de estimação do vetor X (fluxos OD) está fortemente ligado aos relacionamentos lineares descritos pelo sistema da equação (1); Este problema pode ser facilmente formulado como um modelo de Programação Linear. Sabendo que o tráfego em cada enlace y_l é formado pela soma dos fluxos OD que utilizam esse enlace, uma função objetivo para este problema de otimização da estimação do vetor X é dada por:

$$\max \sum_{j=1}^c w_j x_j, \quad (2)$$

onde w_j é um peso para um par OD_{*j*}.

Além da função objetivo, o modelo básico de PL referenciado na literatura para este problema é composto por três conjuntos de restrições:

- restrições de enlace;
- restrições de conservação de fluxo; e
- restrições de positividade.

As restrições de enlace estabelecem que a soma dos fluxos OD que passam pelo enlace y_l não pode ser maior que o valor do tráfego medido nesse enlace. Estas restrições são dadas por:

$$\sum_{j=1}^c A_{lj} x_j \leq y_l \quad l = 1, \dots, r \quad (3)$$

As restrições de conservação do fluxo [Goldscmidt 2000] estabelecem que o tráfego total obtido na saída de um nó é igual à quantidade de tráfego que chega mais a quantidade de tráfego gerada por esse nó, menos a quantidade de tráfego que permanece no nó. Estas restrições são dadas por:

$$\sum y_e - \sum y_s = \sum x_d - \sum x_o \quad (4)$$

onde y_e representa a quantidade de tráfego de entrada de cada enlace, y_s a quantidade de tráfego de saída de cada enlace, x_d representa a quantidade de tráfego destinado ao nó e x_o a quantidade de tráfego originada no nó.

As restrições de positividade são dadas por $x_j \geq 0$, para todo j , significando que nenhum fluxo pode ser negativo.

2.3. Abordagem Bayesiana

A abordagem Bayesiana para a estimação do vetor X , proposta por [Tebaldi e West 1998], calcula a distribuição de probabilidade condicional, $p(X | Y)$, de todos os fluxos OD (representados pelo vetor X) dado o vetor de enlaces Y . Para alcançar esse objetivo é necessário ter a distribuição *a priori* de X , $p(X)$. Tebaldi e West (1998) assumem que $p(X_j)$ segue uma distribuição de Poisson com média λ_j , isto é, $X_j \sim \text{Poisson}(\lambda_j)$. Seja $\Lambda = \{\lambda_1, \dots, \lambda_c\}$ o vetor que representa as taxas médias das distribuições dos elementos do vetor X . Como Λ é desconhecido e precisa ser estimado, torna-se necessário então definir uma distribuição para Λ . Isto nos conduz a um modelo de distribuição conjunta dado por $P(X, \Lambda)$. A idéia é então observar os valores dos enlaces e verificar se eles condicionam o valor de X para obter uma distribuição condicional conjunta dada por $P(X, \Lambda | Y)$.

A obtenção de distribuições *a posteriori*, como $P(X, \Lambda | Y)$, é computacionalmente muito difícil. A abordagem adotada por [Tebaldi e West 1998] para facilitar esse cálculo é aplicar métodos de simulação iterativa como Cadeia de Markov Monte Carlo (MCMC). A simulação da distribuição significa obter um grande número de amostras para representar um histograma completo da distribuição desejada. O objetivo final dessa simulação é calcular $P(X | Y)$.

O mecanismo de simulação iterativo mencionado relaciona a distribuição de probabilidade de X dado Y que se quer obter com a distribuição conjunta que envolve tanto X e Λ pela seguinte equação:

$$p(X | Y) = \sum_{\Lambda} P(X | \Lambda, Y) \quad (5)$$

Assim, o problema consiste em calcular as probabilidades a posterior de $P(\Lambda | X, Y)$ e $p(X | \Lambda, Y)$. O procedimento de simulação começa com uma matriz X^0 que pode ter valores quaisquer e as iterações seguintes são executadas da seguinte forma:

Algoritmo Bayesiano

1. Inicializar Λ^0
2. Calcular X^0 a partir de Λ^0
3. Obter valores de Λ^i a partir de $P(\Lambda | X^i, Y)$
4. Usando este Λ^i , obter valores para o vetor X^{i+1} a partir de $p(X | \Lambda^i, Y)$
5. Repetir os passos 3 e 4 até uma solução viável ser encontrada.

2.4. Abordagem Esperança-Maximização (EM)

Nesta abordagem os pares OD são modelados de acordo com uma distribuição Gaussiana, $X \sim \text{Normal}(\lambda, \Sigma)$, onde os X_j são modelados como variáveis aleatórias normais independentes. Em função disso e considerando a relação da equação (1), a suposição Gaussiana sobre X implica que Y também segue uma distribuição Gaussiana, $Y \sim \text{Normal}(\mathbf{A}\Lambda, \mathbf{A}\Sigma\mathbf{A}')$, onde $\Lambda = \{\lambda_1, \dots, \lambda_c\}$ é o vetor das taxas média dos pares OD e a matriz de covariância é dada por $\Sigma = \phi \text{diag}(\lambda_1^b, \dots, \lambda_c^b)$.

A relação entre a média e a variância da Gaussiana pode sob certas condições ser assumida por $\Sigma_j = \phi \lambda_j^b$, onde ϕ precisa ser estimado juntamente com os λ_j 's.

Seja y_1, \dots, y_k um conjunto de k medições SNMP consecutivas dos enlaces. Este método assume que essas medições correspondem a variáveis aleatórias independentes e identicamente distribuídas [Cao et al 2000]. Seja $\theta = (\Lambda, \phi)$ a equação que representa o conjunto de parâmetros que se quer estimar. A estimativa de máxima verossimilhança é calculada para encontrar o valor máximo da seguinte função:

$$l(\theta | y_1, \dots, y_k) = -\frac{K}{2} \log |A\Sigma A'| - \frac{1}{2} \sum_{k=1}^K (y_k - A\Lambda)' (A\Sigma A')^{-1} (y_k - A\Lambda) \quad (6)$$

O algoritmo EM é usado para calcular uma estimativa para θ , chamada de $\hat{\theta} = (\hat{\Lambda}, \hat{\phi})$. Cao et al (2000) sugerem os seguintes passos:

Algoritmo EM

1. Inicializar θ
2. Cada par OD, índice j , é estimado no tempo t por $X_{j,t} = E[X_{j,t} | \hat{\theta}, Y]$

3. Estimar um novo θ a partir de X .

2.5. Algoritmos Genéticos (AG)

Os algoritmos genéticos (AG) são uma família de modelos computacionais inspirados na evolução, que incorporam uma solução potencial para um problema específico numa estrutura semelhante a de um cromossomo (cadeia de bits que representa uma solução possível para o problema). Os algoritmos AG aplicam operadores de seleção e "cross-over" a essas estruturas de forma a preservar informações críticas relativas à solução do problema [Whitley 1993]. Normalmente, os AG são vistos como otimizadores de funções, embora os AG sejam amplamente aplicados para resolver problemas nos mais diversos campos. De um modo geral, os algoritmos genéticos tentam imitar o princípio de Darwin da seleção natural: um indivíduo adaptado de uma população tende a reproduzir e transmitir seus genes para a próxima geração, tornando as gerações seguintes cada vez melhores. Está provado na literatura que os AG são uma ferramenta eficiente e flexível que pode encontrar a solução ótima ou sub-ótima para problemas lineares ou não-lineares, através da exploração simultânea de várias regiões do espaço. Esta exploração das áreas promissoras é feita exponencialmente através de operadores de seleção, mutação e cruzamento. Diferentemente de outras técnicas de otimização, algoritmos genéticos não exigem condições favoráveis das funções de otimização para que possam ser aplicados [Michalewicz 1996].

Goldberg (1989), um dos precursores dos algoritmos genéticos, descreveu-os como sendo uma pesquisa baseada no mecanismo de seleção e genética natural objetivando a otimização. Os Algoritmos genéticos superam os outros métodos tradicionais de otimização encontrados na literatura por terem as seguintes vantagens:

- a) Trabalham com a codificação dos parâmetros e não com os dados reais;
- b) Fazem busca numa população e não num único ponto;
- c) Usam a informação de aptidão e não outro conhecimento auxiliar; e
- d) Usam regra de transição probabilística e não determinística.

Lutton e Martinez (1994) caracterizaram os algoritmos genéticos como uma técnica estocástica, relativamente lenta, mas de grande eficiência no processo de busca em espaço dimensional elevado.

Whitley (1993) relata que a execução de um algoritmo genético começa com uma população aleatória de cromossomos. Essas estruturas são, então, avaliadas e associadas a uma probabilidade de reprodução de tal forma que as maiores probabilidades são associadas aos cromossomos que representam uma melhor solução para o problema de otimização. A aptidão da solução é tipicamente definida com relação à população atual.

A função objetivo de um problema de otimização é construída a partir dos parâmetros envolvidos no problema. Ela fornece uma medida da proximidade da solução em relação a um conjunto de parâmetros. Estes podem ser incompatíveis, ou seja, se um aumenta o outro diminui. O objetivo é encontrar o ponto de otimalidade. A função objetivo permite o cálculo da aptidão bruta de cada indivíduo, que fornecerá o valor a ser usado para o cálculo de sua probabilidade de ser selecionado para reprodução.

Em [Mitchell 1996], outros conceitos podem ser encontrados sobre os AG: o gene é a representação de cada parâmetro de acordo com o alfabeto utilizado (binário, inteiro ou real). O cromossomo codificado é chamado de fenótipo. A população constitui-se do conjunto de pontos (indivíduos) no Espaço de Busca. Quando existe uma iteração completa do AG que gera uma nova população, temos uma geração. Aptidão bruta é a saída gerada pela função objetivo para um indivíduo da população. Já a aptidão normalizada é a aptidão bruta normalizada, entrada para o algoritmo de seleção. A aptidão máxima será o melhor indivíduo da população corrente e a aptidão média é o meio termo da população corrente.

Deve ser observado que cada cromossomo, chamado de indivíduo no AG, corresponde a um ponto no espaço de soluções do problema de otimização. O processo de solução adotado nos algoritmos genéticos consiste em gerar, através de regras específicas, um grande número de indivíduos, população, de forma a promover uma varredura tão extensa quanto necessária do espaço de soluções.

3. Cenário e Metodologia

Este trabalho foi desenvolvido com base em dois estudos de casos. O primeiro estudo de caso consiste de um cenário proposto por Tebaldi e West (1998) de uma rede com 4 (quatro) nós. O segundo caso consiste de um cenário real da rede GÉANT [Uhlig et al 2006] com 23 nós.

Em ambos os casos foram realizadas 30 rodadas (simulações). No primeiro caso as simulações diferem uma das outras pela quantidade de enlaces utilizados, permanecendo constante o número de nós. No segundo caso, os dados usados nas simulações são diferentes, ou seja, cada simulação usava medições obtidas em momentos diferentes. A implementação computacional do algoritmo genético foi modelada e testada no Matlab [Matlab 1991].

A inicialização das variáveis ocultas, neste caso o vetor X , e dos parâmetros do modelo com valores razoáveis auxilia o algoritmo na convergência para um resultado esperado [Russel e Norvig 2003]. Nessa linha de raciocínio, o modelo (algoritmo) implementa um método de inicialização baseado nos enlaces e nos fluxos OD que se utilizam desses enlaces. O passo principal do procedimento de estimação é calcular todos os parâmetros das distribuições que regem os dados envolvidos na equação (1) para as componentes da matriz de tráfego [Medina et al. 2002]. Neste trabalho, o algoritmo analisa cada coluna da matriz de roteamento (matriz A) à procura dos elementos com valores iguais a 1. A cada valor 1 encontrado na coluna da matriz A , o algoritmo percorre as linhas à procura de incidência de $\{1s\}$, o que indica se um determinado fluxo usa o enlace indicado. Portanto, este método consiste na soma dos valores dos enlaces utilizados por um determinado fluxo dividido pela quantidade de fluxos que passam nesses enlaces.

Uma vez que os parâmetros iniciais são obtidos, o próximo passo é, a partir do Λ inicial (vetor que representa as proporções entre o enlace e todos os seus fluxos passantes), fazer perturbações (mutação) de modo a inferir novos valores de X , seguindo uma distribuição de Poisson. Em cada iteração, é avaliada a resultante da matriz X em função do vetor de tráfego de enlaces Y , buscando sempre a melhor estimação entre todas. Com isso, através dessas avaliações sucessivas dos resultados a estimação pode ou não ser incorporada ao novo conjunto de estimação.

3.1. Estudo de Caso 1 – Topologia com quatro nós

A primeira rede considerada está baseada em [Medina et al 2002] conforme a Figura 1(a). Essa rede é constituída por 4 nós e 7 enlaces. Considera-se inicialmente esse cenário simples pela vantagem de enumerar todos os pares de nós e facilitar a explicação do comportamento do método. Os valores nas conexões dos nós representam as quantidades de tráfego nos enlaces. Cada conexão representa a conectividade entre os pares adjacentes aos PoP (pontos de presença) e cada nó está representando um PoP. Será estimado a quantidade de tráfego em cada fluxo OD, dada a quantidade de tráfego total coletado na entrada e saída de cada nó, respectivamente, oriunda e direcionada para os nós. Os resultados obtidos a partir da estimação proposta são comparados com os valores reais de fluxos OD para cada par, mostrados na Figura 1(b).

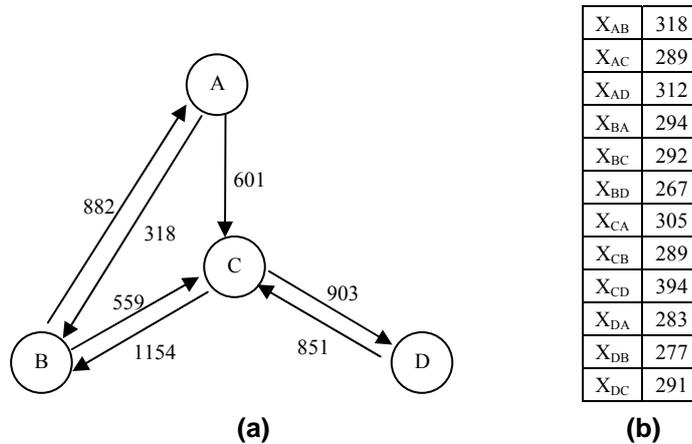


Figura 1: (a) Topologia de quatro nós; (b) Valores reais do tráfego para cada par de fluxo OD

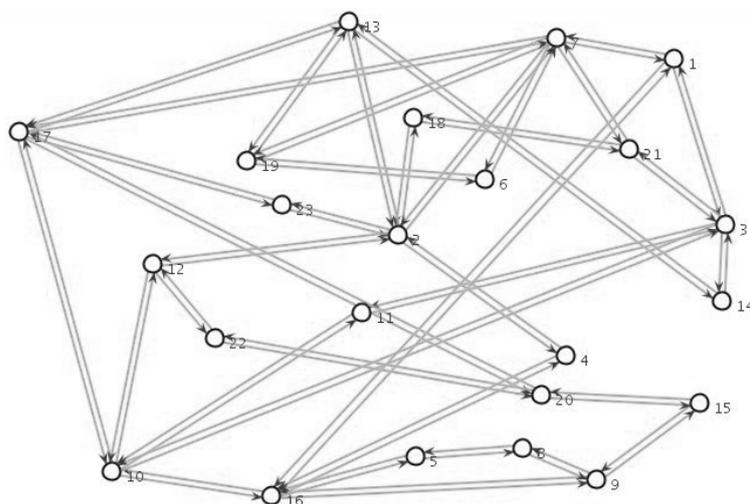
A equação $Y = AX$ compõe, para o caso em estudo, o sistema linear a seguir:

$$\begin{array}{c|c}
 \begin{array}{c} 318 \\ 601 \\ 882 \\ 559 \\ 1154 \\ 903 \\ 851 \end{array} & = & \begin{array}{c} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array} & \times & \begin{array}{c} X_{AB} \\ X_{AC} \\ X_{AD} \\ X_{BA} \\ X_{BC} \\ X_{BD} \\ X_{CA} \\ X_{CB} \\ X_{CD} \\ X_{DA} \\ X_{DB} \\ X_{DC} \end{array}
 \end{array}$$

3.2. Estudo de Caso 2 – Rede GÉANT

A topologia da rede para esse estudo de caso, mostrada na Figura 2(a) está baseada na rede GÉANT [Uhlig et al 2006], uma rede de comunicação de dados multi-gigabit pan-Européia usada para investigação e usos educacionais conectando universidades e instituições de pesquisa. Desenvolver matriz de tráfego exige dados reais para validar os modelos propostos. A rede possui um PoP em cada país europeu. Todos os roteadores são roteadores de borda. É constituída de 23 roteadores interconectados usando 74 enlaces. As rotas intradomínio são obtidas do BGP (Border Gateway Protocol) e a quantidade de tráfego entre fluxos é coletada usando o Netflow. Assim como no estudo de caso 1, os valores nas conexões dos nós representam as quantidades de tráfego nos

enlaces. Cada nó na topologia representa um PoP da GÉANT e cada conexão representa a conectividade entre os pares adjacentes a esses PoP. Utiliza-se o mesmo procedimento de estimação do caso anterior para o segundo cenário. Essa topologia tenta expressar a consistência da convergência uma vez que trata de dados reais. Os resultados obtidos a partir da estimação proposta são comparados com os valores reais da rede. Valores reais de fluxos OD para cada par são mostrados na Figura 2(b).



$X_{(1,2)}$	2276165
$X_{(1,3)}$	13824330
$X_{(1,4)}$	2857357
$X_{(1,5)}$	36580714
$X_{(1,6)}$	1467221
$X_{(1,7)}$	1891468
.	.
.	.
.	.
$X_{(23,17)}$	298672
$X_{(23,18)}$	50251
$X_{(23,19)}$	13904
$X_{(23,20)}$	243006
$X_{(23,21)}$	182364
$X_{(23,22)}$	61980

Figura 2: (a) Topologia usada GÉANT; (b) Exemplo de valores reais do tráfego para cada par de fluxo OD

Essa topologia apresenta pares de nós com enlaces próprios, com ligação direta entre os nós, e pares sem uma ligação direta, ou seja, o tráfego precisa ser roteado através de nós intermediários. Exemplo do primeiro caso é o par (1,3). Existe caminho de '1' para '3' passando por '7' e '21'. Contudo, como há um link direto entre '1' e '3', a rota via '7' e '21' não é utilizada devido o caminho (1,3) ser mais curto. O segundo caso pode ser visto com o tráfego que sai do nó '1' para o nó '2'. Parte do fluxo com 2.276.165 unidades que sai de '1' é destinado ao nó '7' e outra parte ao '2'.

A equação $Y = AX$ está representada a seguir:

$$\begin{array}{c}
 \left| \begin{array}{c}
 16377409 \\
 26983248 \\
 104787085 \\
 80597391 \\
 28952052 \\
 92348785 \\
 \vdots \\
 \vdots \\
 8405377 \\
 53991705 \\
 16160816 \\
 9519418 \\
 31804336 \\
 26394286
 \end{array} \right| = \left| \begin{array}{cccccccc}
 0 & 1 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 1 & 0 & 1 & 0 & 1 & 1 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 1 \\
 \vdots & \vdots \\
 \vdots & \vdots \\
 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 1 & 1 & 0 & 1 & 1 \\
 0 & 0 & 0 & 0 & 0 & 0 & \dots & 1 & 0 & 0 & 1 & 0 & 0
 \end{array} \right| \times \left| \begin{array}{c}
 X_{(1,2)} \\
 X_{(1,3)} \\
 X_{(1,4)} \\
 X_{(1,5)} \\
 X_{(1,6)} \\
 X_{(1,7)} \\
 \vdots \\
 \vdots \\
 X_{(23,17)} \\
 X_{(23,18)} \\
 X_{(23,19)} \\
 X_{(23,20)} \\
 X_{(23,21)} \\
 X_{(23,22)}
 \end{array} \right|
 \end{array}$$

Há um número infinito de soluções para o vetor X , pois existem mais incógnitas do que equações. O sistema linear gerado pelo problema foi resolvido neste trabalho através de algoritmo genético. Correspondem a 506 fluxos OD estimados.

4. Resultados e Discussão

Os resultados obtidos do processamento de cada um dos casos descritos estão apresentados a seguir:

No caso 1, tabela 1 tem-se os resultados da estimação usando Programação Linear (PL), inferência Bayesiana e Esperança-Maximização (EM), para a mesma topologia de rede e os mesmos dados de entrada, obtidos e apresentados em [Medina et al 2002]. Para PL e inferência Bayesiana, usou-se uma MT Poisson, e para o método EM foi usado MT Gaussiana. Os pesos foram fixados em 1 para cada enlace e as rotas foram computadas de acordo com o número de salto mais curto.

Os resultados obtidos neste trabalho com algoritmo genético foram acrescidos e apresentados à tabela, comparativamente com os valores reais do tráfego da rede e, de uma maneira geral, com os demais métodos. A tabela mostra a MT original, o valor estimado para cada par OD e a coluna ERRO mostra a diferença percentual entre os resultados estimados pelo modelo e os valores reais de cada método.

A partir destes resultados pode-se avaliar a precisão das soluções obtidas para as diversas combinações de conjuntos de restrições e função objetivo utilizadas. O erro médio foi de 98,2% para o método PL, 12,9% para o método Bayesiano e 7,4% para o método EM. Com AG o erro médio foi de 3,9%. O método PL é claramente pior do que os demais. Pode-se ver que este método usa valores zero para alguns pares OD. Com isso temos várias estimações de erro 100% ou mais o que gera uma taxa média muito alta. No caso dos métodos Bayesianos e EM, mais ou menos, metade das estimativas são excessivas em suas previsões e a outra metade é de estimativa baixa. Entre esses dois métodos, EM apresenta um ganho em relação a Bayesiana.

Com AG, foram estimados valores com variação menor, ou seja, a taxa de erro girou entre 0% e 6,7%, o que possibilitou uma taxa média de erro baixa. Com isso, vê-se que a estimação apresenta o equilíbrio, independente de fluxos que compartilham ligações ou não, diferentemente do caso de inferência Bayesiana, que apresenta seu pior erro para o par OD CB, ou seja, sua estimativa de erro pode ser fortemente correlacionada com ligações compartilhadas.

Tabela 1. Estimação para Topologia de quatro nós

Original MT (Poisson)	PL		Bayesiana		Original MT (Gaussiana)	EM		AG	
	MT Estimada	Erro (%)	MT Estimada	Erro (%)		MT Estimada	Erro (%)	MT Estimada	Erro (%)
AB: 318	318	0	318	0	318,65	318,65	0	317	0,3
AC: 289	601	107	342	18	329,48	289,98	13	277	4,2
AD: 312	0	100	259	17	277,18	318,36	15	322	3,2
BA: 294	579	96	334	14	298,14	298,14	0	279	5,1
BC: 292	559	91	310	6	354,81	360,97	1,6	280	4,1
BD: 267	0	100	249	7	355,39	347,94	2	279	4,5
CA: 305	303	0,6	291	5	327,20	317,34	3	301	1,3
CB: 289	0	100	361	25	330,04	373,65	13	274	5,2
CD: 324	903	178	395	22	253,01	217,32	14	303	6,5
DA: 283	0	100	257	9	320,50	329,07	3	302	6,7
DB: 277	851	207	245	12	291,52	246,60	15	277	0
DC: 291	0	100	349	20	310,40	344,82	11	272	6,5
	Erro Médio:	98,2	Erro Médio:	12,9		Erro Médio:	7,4	Erro Médio:	3,9

Os resultados do uso de AG na topologia do Caso 1 são apresentados na Figura 3. O gráfico mostra o percentual de erro médio sobre o número de vezes que o método encontra uma taxa de erro de Y inferior a uma taxa de erro anterior. Nessa simulação, fora estipulado um valor inicial de 3% para taxa de erro da matriz Y. Cada valor médio inferior a este é tomado como o valor mais baixo e assim sucessivamente. Em relação à taxa de erro médio de estimação de X, não é possível simulá-la controladamente como em Y, uma vez que se trata do objetivo do método. Vimos nesse caso o bom desempenho da taxa de erro de X, encerrando a simulação com a taxa de 3,9%, uma melhora de 46% em torno da taxa de erro inicial para a matriz X.

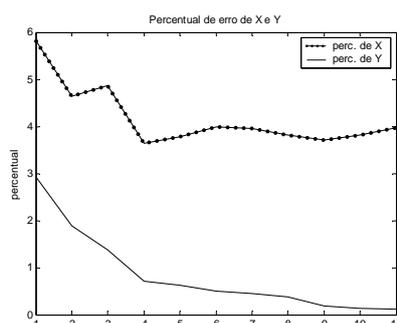


Figura 3: Percentual de erro de estimação da Matriz X e Matriz Y - Topologia de quatro nós

No caso 2, tabela 2 tem-se os resultados da estimação usando algoritmo genético para rede GÉANT. Estes resultados são comparados com os valores reais do tráfego da rede. A coluna ERRO, da mesma forma, mostra a diferença percentual entre os resultados estimados pelo modelo e os valores reais. A ilustração dos dados fora suprimida por se tratar de um expressivo número de fluxo OD contido nessa rede. Diferentemente dos resultados apresentados no caso 1, a aplicação do resultado na rede GÉANT trouxe valores altos de taxa de erro o que gerou um valor médio acima de 60%. O erro individual variou de 1,1% até 295%. Ao ser considerado dados reais podem existir fluxos com valor zero para um fluxo OD, indicando que no momento de coleta dos dados não havia fluxo entre um determinado par. Esses casos são tratados pelo AG proposto.

Tabela 2. Estimação para rede GÉANT

Dados Reais	AG	
	Dados Estimados	Erro
$X_{(1,2)}$: 2276165	1612244	70,8
$X_{(1,3)}$: 13824330	6471120	46,8
$X_{(1,4)}$: 2857357	5302016	185,6
$X_{(1,5)}$: 36580714	17200493	47,0
$X_{(1,6)}$: 1467221	67913	4,6
$X_{(1,7)}$: 1891468	1303698	68,9
.	.	.
.	.	.
.	.	.
$X_{(23,17)}$: 298672	588862	197,2
$X_{(23,18)}$: 50251	78492	156,2
$X_{(23,19)}$: 13904	18395	132,3
$X_{(23,20)}$: 243006	457337	188,2
$X_{(23,21)}$: 182364	267528	146,7
$X_{(23,22)}$: 61980	91421	147,5
	Erro Médio:	62,8

5. Conclusões e Trabalhos Futuros

O uso de algoritmos genéticos para estimação de matrizes de tráfego OD constitui-se algo inédito nesta área de pesquisa. Este trabalho utiliza AG com o objetivo de melhorar a precisão até então obtida por outros métodos. Os resultados obtidos quando comparados com aqueles de Medina et al (2002) e com a mesma topologia de rede e os mesmos dados de entrada mostram que há uma melhora significativa na estimação dos fluxos OD.

Resultados obtidos em trabalhos anteriores apresentam erros médios de estimação da ordem de 98,2% para PL, 12,9 % para abordagem Bayesiana e 7,4% para o algoritmo EM. Trabalhos mais recentes mostram que a programação linear e a adição de novas restrições [Silva et al. 2007] proporcionam resultados significativamente superiores aos anteriormente citados. Contudo, o esforço de estabelecer restrições para redes com mais de dez nós e com o dobro de enlaces e fluxos OD não estimulam a sua utilização.

Este trabalho conseguiu obter erros médios de estimação na ordem de 3,9%, em alguns testes, para uma topologia de rede com quatro nós, sete enlaces e doze fluxos OD. Foram realizados testes com um número maior e menor de enlaces para a mesma quantidade de nós e os resultados estão na mesma ordem.

Os resultados obtidos no segundo estudo de caso, com uma rede de 23 nós (GEANT), mostram que o percentual de erro médio é muito grande (62,8%), desestimulando a utilização do AG. Contudo, o AG aplicado neste trabalho ainda é incipiente, pois usa somente o operador de mutação. Planeja-se tentar heurísticas mais elaboradas e espera-se que esta abordagem, com as implementações dos demais operadores, venha mostrar resultados significativamente melhores.

Os resultados obtidos até o momento estimulam a continuação nessa direção. Trabalhos futuros planejados são: i) teste do algoritmo com outras topologias de redes de backbone, como a RNP (Rede Nacional de Pesquisa), cujos os dados já estão sendo armazenados; ii) uso de novas meta-heurísticas e comparação dos resultados.

6. Agradecimentos

Este trabalho foi desenvolvido com o apoio da Funcap (Fundação Cearense de Apoio a Desenvolvimento Científico e Tecnológico).

Referências

- Cao J., Davis D., Weil S. V., and B. Yu. (2000). Time-Varying Network Tomography. *J. of the American Statistical Association*.
- Goldberg, D.E. (1989). Genetic Algorithms in Search, Optimization and Machine Learning. Editora: Addison-Wesley. New York.
- Goldschmidt, O. (2000). ISP Backbone Traffic Inference Methods to Support Traffic Engineering. Em Internet Statistic and Metrics Analysis (ISMA) Workshop, San Diego, CA.
- Holland, J.H. (1975). Adaptation in Natural and Artificial Systems. The University of Michigan Press, Ann Arbor, 2nd edition.

- Lutton, E.; Martinez, P. (1994). A Genetic Algorithm for the Detection of 2D Geometric Primitives in Images; 1051-4651/94 IEEE. INRIA - Rocquencourt, B.P. 105, 78153 LE CHESNAY Cedex, France.
- MATLAB for Windows User's Guide, The Math Works Inc., 1991.
- Medina, A., Taft, N. Salamatain, K. Bhattacharyya, S., Diot, C. (2002). Traffic Matrix Estimation: Existing Techniques and New Directions. SIGCOMM.
- Michalewicz, Z. (1996). Genetic Algorithms + Data Structures = Evolution Programs, 3th edition, Springer - Verlag, New York.
- Mitchell M. (1996). An Introduction to Genetic Algorithms. MIT Press.
- Rschendorf L. (1995). Convergence of the iterative proportional fitting procedure. Annals of Statistics, pages 1160–1174.
- Russell , Norvig. (2003). Artificial Intelligence: A Modern Approach. Prentice Hall.
- Silva, A. N., Lima, H. O., Pereira, S.S.L., Nobre, D. N., Silva, J. L. C., Maia, J. E. B., Cunha, P. R. F. (2007). Estimação de Matrizes de Tráfego de Backbones utilizando Restrições de Proporcionalidade em Modelo de Programação Linear. SBRC.
- Soule, A., Lakhina, A., Taft, N., Papagiannaki, K., Salamatian, K., Nucci, A., Crovella, M., e Diot, C. (2005). Traffic matrices: Balancing measurements, inference and modeling. *In Proc. of the ACM SIGMETRICS'05*, Banff, Canada.
- Soule, A., Nucci, A., Cruz, R., Leonardi, E., e Taft, N. (2004). How to identify and estimate the largest traffic matrix elements in a dynamic environment. *In Proc. of the ACM SIGMETRICS'04*, Nova Iorque, NY, EUA.
- Strang G. (1993). Introduction to Linear Algebra. Wellesley-Cambridge Press.
- Tan, K. C., Li, Y., Murray-Smith, D .J. and Sharman, K. C. (1995). Nonlinear Parameter Estimation via the Genetic Algorithm, Proc. IEE/IEEE Int. Conf. On GA in Eng. Syst.: Innovations and Applications, Sheffield, U.K., pp. 164-169.
- Tebaldi, C e West, M. (1998). Bayesian Inference of Network Traffic Using Link Count Data. *Journal of the American Statistical Association*, pag. 557-573.
- Teixeira, R., Duffield, N., Rexford, J., e Roughan, M. (2005). Traffic matrix reloaded: Impact of routing changes. *In Proc. of the Passive and Active Measurement Workshop-PAM'2005*, Boston, MA, EUA.
- Uhlig, S. and Quoitin, B. and Balon, S. and Lepropre, J. (2006). Providing Public Intradomain Traffic Matrices to the Research Community. SIGCOMM Comput. Commun. Rev. ACM Press.
- Vardi, Y. (1996). Network tomography: Estimating source-destination traffic intensities from link data, *Journal of the American Statistical Association*, 91, 365-377.
- Whitley D. (1993). A Genetic Algorithm Tutorial. Computer Science Department, Colorado State University.
- Ying-ping Chen. (2004). Extending the Scalability of Linkage Learning Genetic Algorithms: Theory and Practice. Dissertation, University of Illinois at Urbana-Champaign.

Zhang, Y., Roughan, M., Duffield, N., e Greenberg, A. (2003a). Fast accurate computation of large-scale IP traffic matrices from link loads. *In Proc. of the ACM SIGMETRICS'03*, San Diego, CA, EUA.

Zhang, Y., Roughan, M., Lund, C., e Donoho, D. (2003b). An information-theoretic approach to traffic matrix estimation. *In Proc. of the ACM SIGCOMM'2003*, Karlsruhe, Germany.