

# Coleta e Análise de Características de Fluxo para Classificação de Tráfego em Redes Definidas por Software

---

**Rodolfo Vebber Bisol , Anderson Santos da Silva, Cristian Cleder Machado,  
Lisandro Zambenedetti Granville, Alberto Schaeffer-Filho**

*Universidade Federal do Rio Grande do Sul (UFRGS), Brazil*

XXXIV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos (SBRC 2016)

30 de maio a 03 de junho 2016

Salvador – Bahia - Brasil

# Roteiro

- ① Introdução
- ② Conceitos básicos
- ③ Arquitetura para seleção de características
- ④ Experimentos
- ⑤ Conclusão

# Introdução

- Classificação de tráfego
  - Conjunto de técnicas para encontrar padrões de tráfego
  - Aplicações: detecção de ataques maliciosos, modelagem de tráfego
  - Desejável técnicas com boa acurácia de classificação
- Caracterizar tráfego em redes tradicionais possui desafios
  - Dispositivos heterogêneos (switches proprietários)
  - Controle de rede distribuído nos dispositivos encaminhadores de pacote
- Software-Defined Networking (SDN)
  - Separa o plano de controle e os dispositivos encaminhadores
  - Facilita a inserção de software na rede
  - Precisa de suporte para classificação de tráfego
    - Contadores de pacotes, por exemplo, não oferece muita informação

# Introdução

- Trabalho anterior
  - Arquitetura para identificar, estender, e selecionar um conjunto de características de fluxo (Silva 2015, Feature Selection SDN)
    - Derivada dos contadores nativos do protocol OpenFlow
    - Não explora algoritmos não-supervisionados para classificação
    - Tempo de execução pode ser muito alto com o Algoritmo Genético
- Contribuição deste trabalho
  - Explora novas técnicas para encontrar o conjunto ótimo de características para diferentes tipos de fluxos
    - Sequential Backward Selection(SBS)
    - K-means

# Conceitos básicos

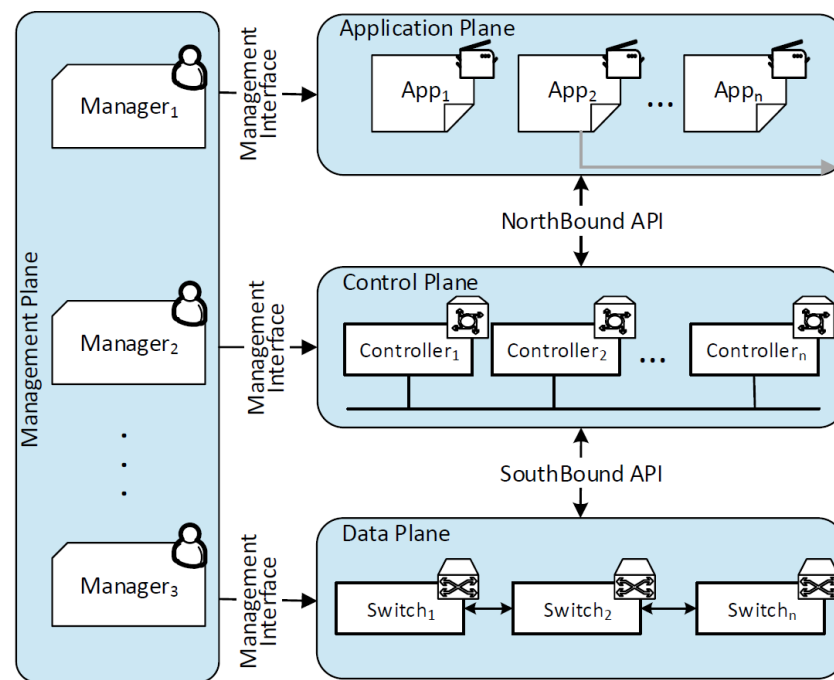
- Técnicas de Detecção e Classificação
  - Capazes de identificar padrões dentro do tráfego na rede
  - Estratégias baseadas em número de porta e inspeção de payload não são confiáveis e podem ser custosas
  - Machine learning é uma tendência promissora em relação a isso
  - Aprendizado Supervisionado
    - Adequada para conjunto de ataques conhecidos
    - Não lida com novos tipos de ataques
    - Support Vector Machines (SVM) possui boa acurácia
  - Aprendizado não supervisionado
    - Adequado para detector novos tipos de ataques
    - Precisa a interação humana para determinar o rótulo
    - K-means possui bons resultados e um tempo aceitável

# Conceitos básicos

- Seleção de características
  - Número excessivo de características para descrever um fluxo
  - Características irrelevantes ou redundantes devem ser removidas
- A precisão da classificação depende das características
  - Poucas: não há informação suficiente para classificações precisas
  - Demasiadas: ruído causado pelo grande número de variáveis e custo computacional
- Técnicas principais
  - Principal Component Analysis (PCA)
  - Genetic Algorithms (GA)
  - Sequential Backward Selection (SBS)

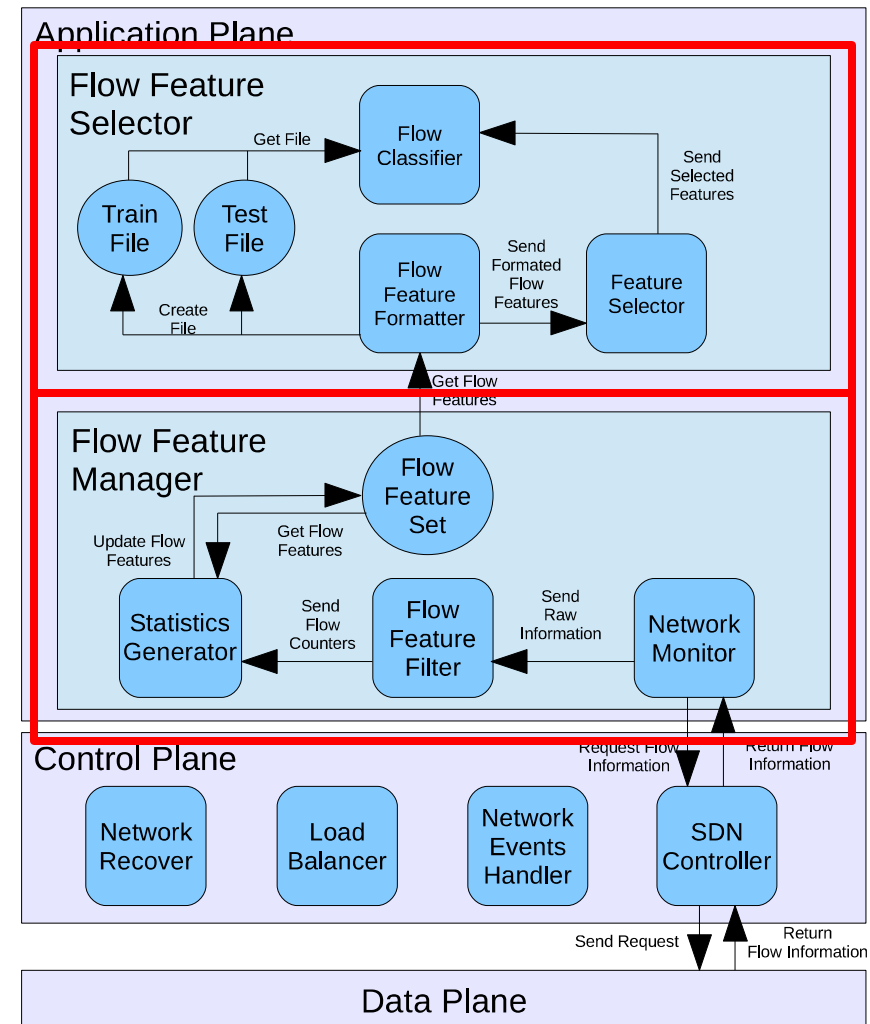
# Conceitos básicos

- Software-Defined Networking(SDN)
  - Arquitetura modular
  - Lógica de controle centralizada
  - protocolo OpenFlow
  - Southbound API
- Os dispositivos de rede tornam-se simples encaminhadores de pacotes
- Características de fluxo nativas são limitadas no protocolo OpenFlow



# Arquitetura para Flow Feature Selection

- Principais funcionalidades
  - Conjunto estendido de características de fluxo
  - Capacidade de encontrar o conjunto ótimo de características
- Arquitetura proposta inclui dois componentes
  - Flow Feature Manager
    - Busca informação na rede, coleta estatísticas, e computa características de fluxo adicionais
  - Flow Feature Selector
    - Seleciona as melhores características de fluxo para classificação de tráfego



# Flow Features extendidas

- Limitações OpenFlow
  - Contadores não permitem a detecção de comportamento de tráfego atípico, como rajadas de pacotes
- Extensão dos contadores nativos do OpenFlow através da análise de estatísticas
  - Os três contadores básicos, `byte_count`, `packet_count` e `duration` foram estendidos para criar um conjunto de características de fluxo mais preciso
  - Produzir mais informações descritivas sobre o comportamento do tráfego, tais como bytes por segundo, média de pacotes por segundo e variância

# Flow Features extendidas

Características estadísticas	Características escalares	Características complejas
Bytes per second mean	Bytes per second maximum value	Packet inter-arrival-time Fourier Transform 1 <sup>st</sup> Component
Bytes per second variance	Bytes per second minimum value	Packet inter-arrival-time Fourier Transform 2 <sup>nd</sup> Component
Packets per second mean	Packets per second maximum value	Packet inter-arrival-time Fourier Transform 3 <sup>rd</sup> Component
Packets per second variance	Packets per second minimum value	Packet inter-arrival-time Fourier Transform 4 <sup>th</sup> Component
Packets length mean	Packets length maximum value	Packet inter-arrival-time Fourier Transform 5 <sup>th</sup> Component
Packets length variance	Packets length minimum value	Packet inter-arrival-time Fourier Transform 6 <sup>th</sup> Component
Packets length 1 <sup>st</sup> quartiles	Packets inter-arrival-time maximum Value	Packet inter-arrival-time Fourier Transform 7 <sup>th</sup> Component
Packets length 3 <sup>rd</sup> quartiles	Packets inter-arrival-time minimum value	Packet inter-arrival-time Fourier Transform 8 <sup>th</sup> Component
Packet inter-arrival-time mean	Flow duration	Packet inter-arrival-time Fourier Transform 9 <sup>th</sup> Component
Packet inter-arrival-time variance	Flow size in packets	Packet inter-arrival-time Fourier Transform 10 <sup>th</sup> Component
Packet inter-arrival-time 1 <sup>st</sup> quartiles	Flow size in bytes	
Packet inter-arrival-time 3 <sup>rd</sup> quartiles		

# Flow Features extendidas

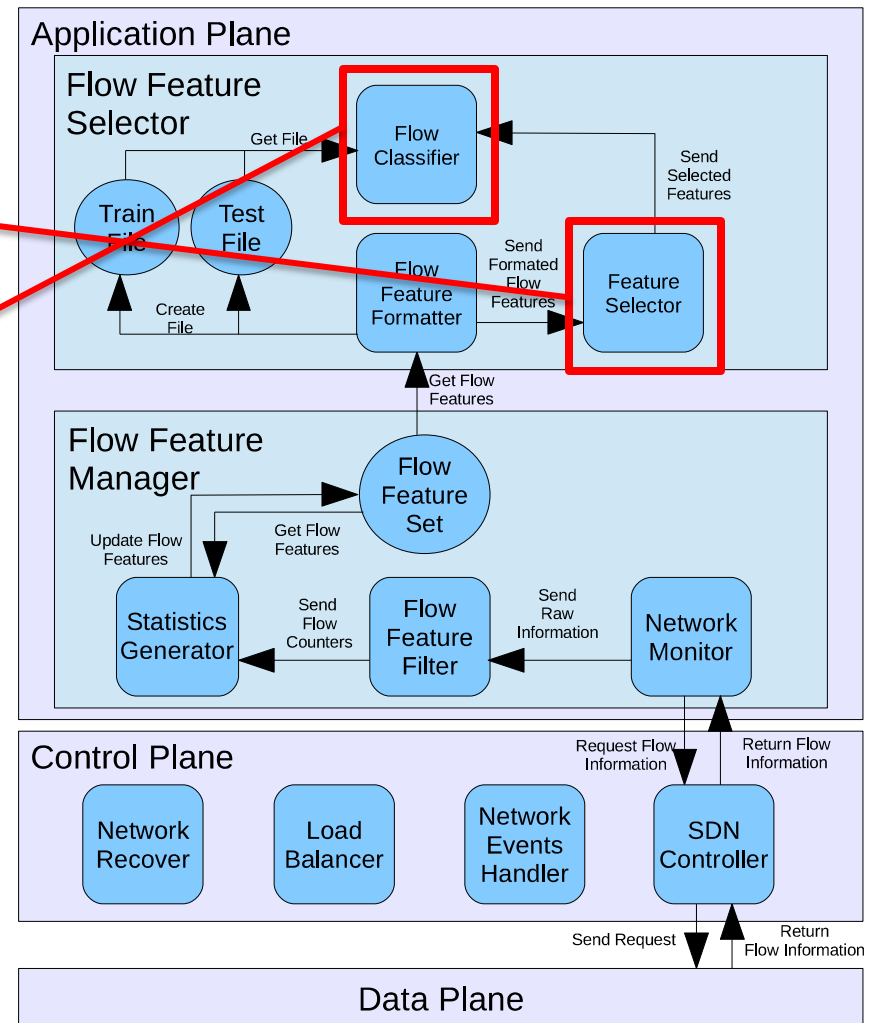
- Conjunto estendido de características de fluxo
  - Características estatísticas
    - Média, variância, primeiro e terceiro quartis calculado sobre as características nativas do OpenFlow
    - Melhores descritores em comparação com os contadores de pacotes nativos
  - Características Escalares
    - Convenientes para indicar perfil instantâneo dos fluxos de tráfego, como a duração
    - Importante na detecção de comunicações de vida curta ou rajadas de pacotes, uma vez que os recursos estatísticos por si só não são sensíveis a estes perfis de tráfego
  - Características complexas
    - Transformada de Fourier discreta (DFT) de pacotes inter-hora de chegada
    - Oferecem informação sobre o comportamento do pacote

# Flow Features estendidas

- Arquitetura estendida para seleção de características
  - Algoritmos antigos
    - Classificação: SVM
    - Seleção de características: PCA e GA
  - Sequential Backward Selection (SBS) para Seleção de Características
    - Escolhe uma característica a ser eliminada consiste partindo de um subconjunto
    - Avalia a qualidade deste subconjunto com cada característica eliminada uma a uma
    - Termina quando nenhuma eliminação melhora a acurácia
  - K-means para classificação
    - Algoritmo não-supervisionado
    - Agrupa dados por similaridade, como por exemplo, distância euclidiana
    - Ideal para classificar dados sem conhecimento *a priori*
    - A medida de silhueta ajuda a verificar se cada grupo gerado é homogêneo
    - **Silhueta: boa forma de quantificar a homogeneidade dos clusters gerados**

# Experimentos

- Extensão de arquitetura
  - Sequential Backward Selection (SBS) para realizar seleção de características de fluxo juntamente com PCA e GA já implementados
  - K-means para classificação de tráfego usando análise de sua silhueta juntamente com SVM



# Experimentos

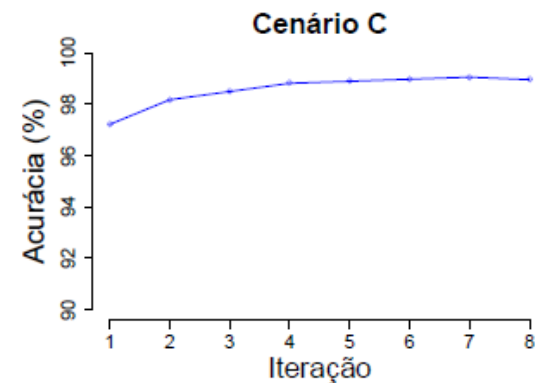
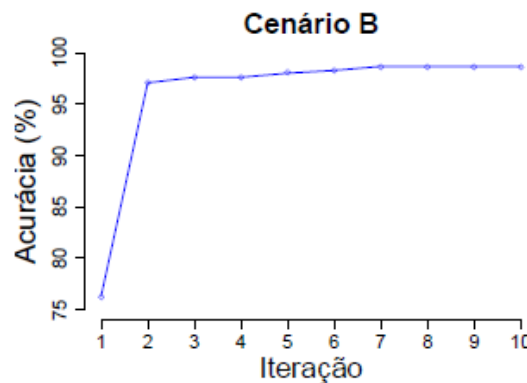
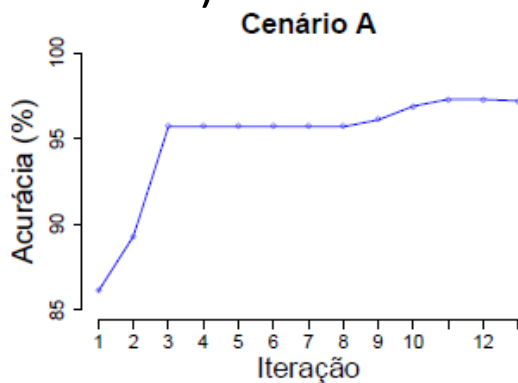
- Avaliação dos experimentos

- Medir a acurácia de classificação
  - Usando subgrupos de características de tráfego descobertas através de algoritmos como Genetic Algorithm (GA), Principal Component Analysis (PCA), SBS e k-means
- Topologia com 21 switches e 64 hosts
- Emulador Mininet e controlador Floodlight
- Definimos 4 tipos de tráfego
  - DDoS attack
  - FTP traffic
  - Video streaming using VLC Media Player
  - Background traffic generated using Scapy

Cenário	Perfil de tráfego
Stream (A)	Ataques DDoS (10%), Tráfego FTP (10%), Stream de video (50%) e background (30%)
Mix (B)	Ataques DDoS (30%), Tráfego FTP (10%), Stream de video (40%) e background (20%)
DDoS (C)	Ataques DDoS (50%), Tráfego FTP (7.5%), Stream de video (25%) e background (17.5%)

# Experimentos

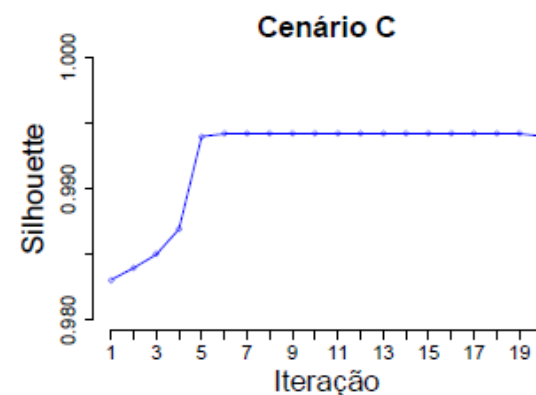
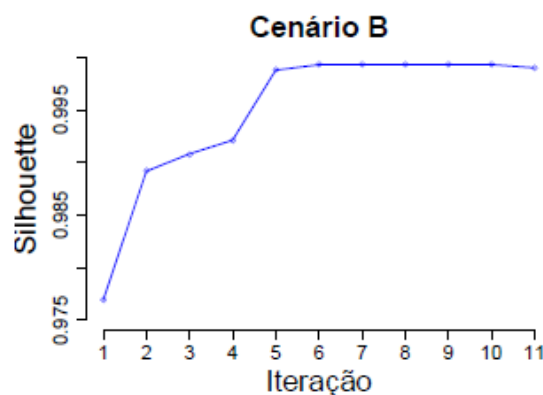
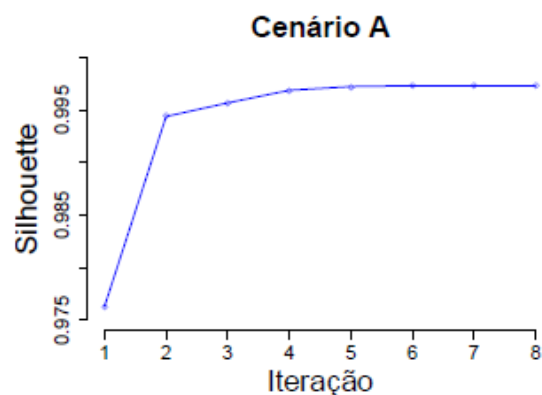
- Acurácia de classificação obtida com o algoritmo SBS
  - a última iteração resulta em um decréscimo da acurácia. Isto indica o momento em que o algoritmo interrompe a execução
  - existem momentos em que a acurácia não sofre alteração (cenários A e B).



	Todas Características	SBS	PCA	GA
Cenário A	86%	97.2%	94.8%	97.2%
Cenário B	76.26%	98.53%	96.93%	98.67%
Cenário C	97.2%	99.04%	99.04%	99.12%

# Experimental Evaluation

- Evolução da Silhouette a cada iteração do algoritmo SBS



	Todas Características	SBS	PCA	GA
Cenário A	0.976295	0.997305	0.997305	0.983541
Cenário B	0.976893	0.999317	0.999092	0.991594
Cenário C	0.983061	0.994187	0.988336	0.983100

# Experimentos

- Clusterização realizada pelo K-means para os cenários A, B e C, respectivamente
  - São comparados os resultados do conjunto de características completo (à esquerda) com os resultados dos subconjuntos criados por cada um dos algoritmos de seleção de características

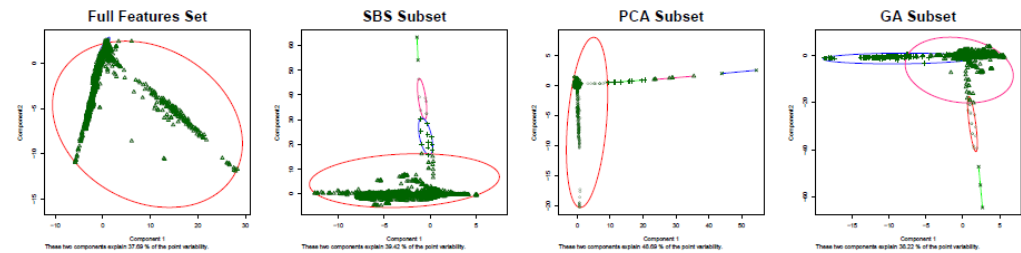


Figura 4. Resultado da clusterização para o cenário A.

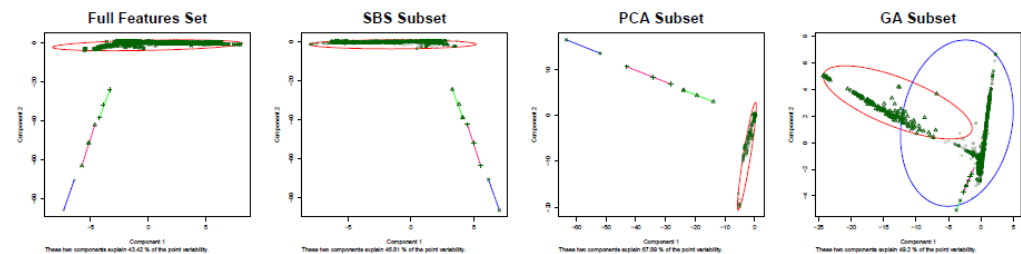


Figura 5. Resultado da clusterização para o cenário B.

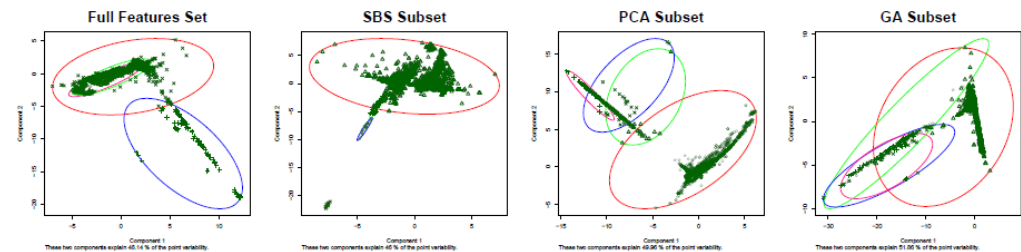
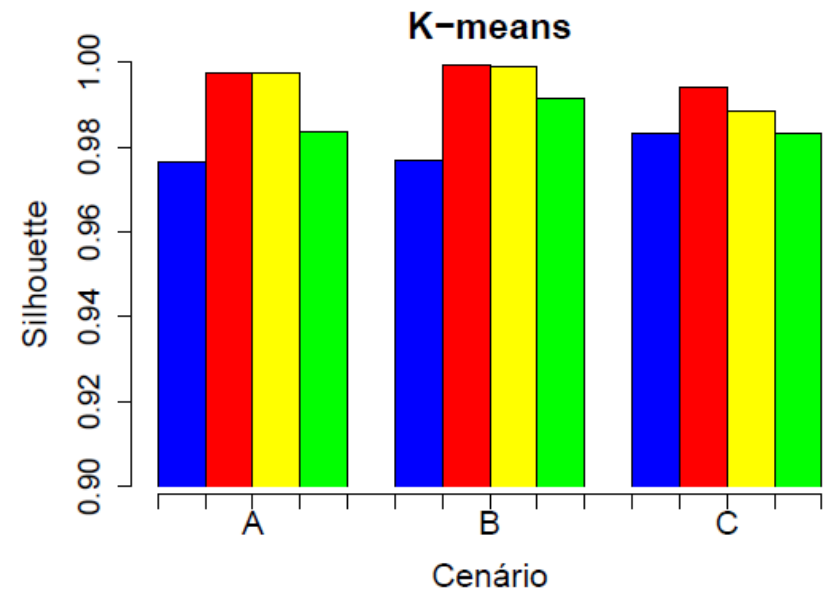
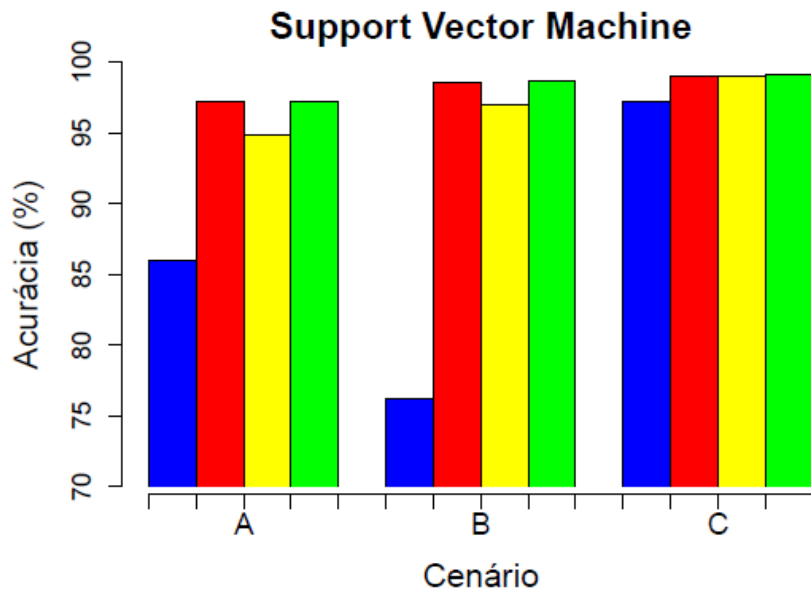


Figura 6. Resultado da clusterização para o cenário C.

# Experimentos

- Comparação com resultados anteriores



■ All features  
 ■ SBS  
 ■ PCA  
 ■ GA  
 ■ All features  
 ■ SBS  
 ■ PCA  
 ■ GA

	SVM			K-means		
	SBS	PCA	GA	SBS	PCA	GA
Cenário A	21min 42s	1min 51s	16h 36min	22min 10s	3min 13s	35h 42min
Cenário B	19min 34s	2min 07s	15h 40min	26min 14s	2min 59s	29h 42min
Cenário C	12min 38s	1min 47s	11h 11min	42min 57s	3min 06s	35h 59min

# Conclusão

- Extensão de arquitetura
  - Para seleção de características de fluxo de forma acurada
  - Conjunto ótimo de características de fluxo
    - Usando o algoritmo SBS e K-means
- Resultados experimentais indicam que
  - SBS tem um desempenho muito bom por apresentar uma grande melhoria na qualidade da classificação do SVM e na clusterização do K-means e também um tempo de execução muito inferior ao algoritmo genético sem perda de qualidade.
- Trabalhos Futuros
  - utilização de tráfego real e não apenas tráfego sinteticamente criado.
  - coleta de características com Deep Packet Inspection (DPI)

# Coleta e Análise de Características de Fluxo para Classificação de Tráfego em Redes Definidas por Software

---

**Anderson Santos da Silva**

*Universidade Federal do Rio Grande do Sul (UFRGS), Brasil*

*assilva@inf.ufrgs.br*

**Obrigado!**

SBRC 2016 XXXIV Simpósio Brasileiro de Redes de Computadores e Sistemas Distribuídos  
30 de maio a 03 de junho 2016  
Salvador – Bahia - Brasil