A Visual Evaluation Study of Graph Sampling Techniques

Evaristo Calisto Nhassengo

INSTITUTE OF MATHEMATICS AND COMPUTER SCIENCES

Outubro, 2018

Cronograma

Cronograma

Introdução

Nas ultimas décadas a analise e Visualização tem evoluído de forma muito rápida, com aplicações em redes sociais, segurança, computação de alto desempenho, etc. No entanto, à medida que o tamanho de um gráfico cresce, as tarefas de analisar e visualizar se tornam extremamente difíceis.

Objectivos

- Objectivo Geral:
  1. Fazer um estudo de avaliação visual de técnicas de amostragem sobre grafos.
- Objectivos específicos:
  1. Estudar a influencia das técnicas de amostragem nas propriedades estatísticas do grafo;
  2. Estudar a influencia das técnicas de amostragem na visualização do grafo;

Porque amostragem em grafos

A amostragem em grafos é necessária na análise de grafos por várias razões:

1. Exibir até mesmo um gráfico relativamente pequeno de vários milhares de vértices em uma tela é desafiador devido ao limite no tamanho da tela.
2. A segunda razão é que a análise de um gráfico grande é cara.
3. Grafos incompletos.

Pelas razoes acima, a amostragem em grafos tem como objectivo reduzir a complexidade do desenho do grafo , preservando as propriedades do grafo original.

Métodos de amostragem em grafos

Selecionam aleatoriamente vértices e nós, e formam um subgrafo que represente o grafo original.

- **Node Sampling** Os vértices são amostrados aleatoriamente e uniformemente. E é criado um subgrafo dos vértices amostrados e as arestas existentes no grafo original.
  - **Random Degree Node (RDN)**
- **Edge Sampling** Arestas são amostradas aleatoriamente e uniformemente.
  - **Totally induced edge sampling**
  - **Ramdom Node-Edge (RNE)**
- **Traversal-based sampling**
  - **breadth-first sampling**
  - **Random-first sampling**
  - **Snowball sampling**
  - **Random walk sampling**

Base de Dados de Grafos

As Redes sociais , grafos de citação, gráficos de comunicação por e-mail e grafos da Internet são baixados da Stanford Network Analysis Platform (SNAP) `http://snap.stanford.edu/data/index.html` . Os garfos mundo pequeno e os grafos aleatórios são criados a partir do NetworkX.

Base de Dados de Grafos

Dataset	Graph Type	Model	#Vertices	#Edges
Random	Directed	Model	10,000	100,246
Small-World	Undirected	Model	10,000	21,895
Scale-Free	Directed	Model	10,000	18,838
Email	Directed	Real	265,214	420,045
Citation	Directed	Real	34,546	421,578
Internet	Directed	Real	10,876	39,994
Facebook	Undirected	Real	4,039	88,234
U.S. Flight	Undirected	Real	235	1,297

Figura: Os conjuntos de dados e suas propriedades

Comparações estatísticas

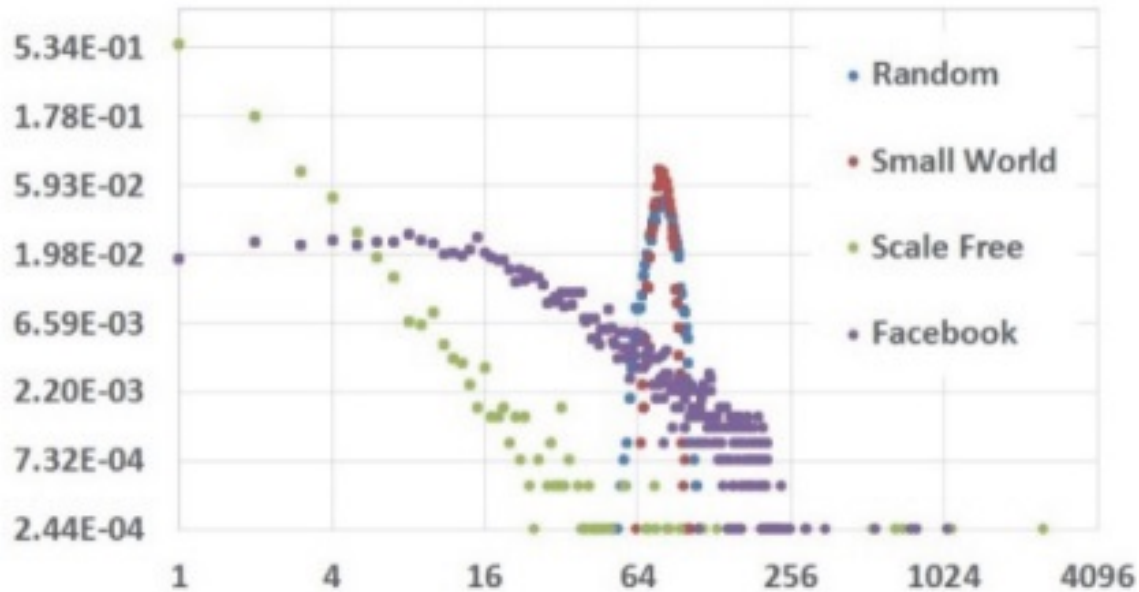

Figura: Distribuição de graus do modelo de gráfico aleatório (azul), mundo pequeno (vermelho), livre de escala (verde) e grafo de rede social (magenta)

- Grafo de rede social é um grafo complexo, diferente que qual quer outro modelo de grafo teórico

Comparações estatísticas

Propriedades dos grafos usadas para comparação dos grafos.

1. DD- Degree distribution
2. ANDD- Average neighbor degree distribution
3. DCD- Degree centrality distribution
4. NBCD-Node betweenness centrality distribution
5. EBCD- Edge betweenness centrality distribution
6. LCCD- Local clustering coefficient distribution
7. EVCD- Eigenvector centrality distribution
8. InDD- In-degree distribution
9. OutDD- Out-degree distribution

Comparações estatísticas

As técnicas de amostragem são avaliadas baseadas na comparação das distribuições propriedade dos grafos entre os métodos de amostragem.
Um bom método de amostragem deve produzir um grafo com resultados de amostragem próximos ao grafo original.
$\Longrightarrow$
As distribuições de probabilidade das propriedades dos dois grafos devem ter uma curta distância entre elas.

- Skew divergence (SD)

Comparações Visuais

Para comparar visualmente os métodos de amostragem foi usado o Gephi.

1. Desenhar o garfo original, e usar este layout para para todos os grafos amostrados.
   - Preserve a localização do vértice;
   - Preserve a cor e tamanho;
   - Não são preservados os atributos das arestas (Cor, Pesos).

Comparações estatísticas

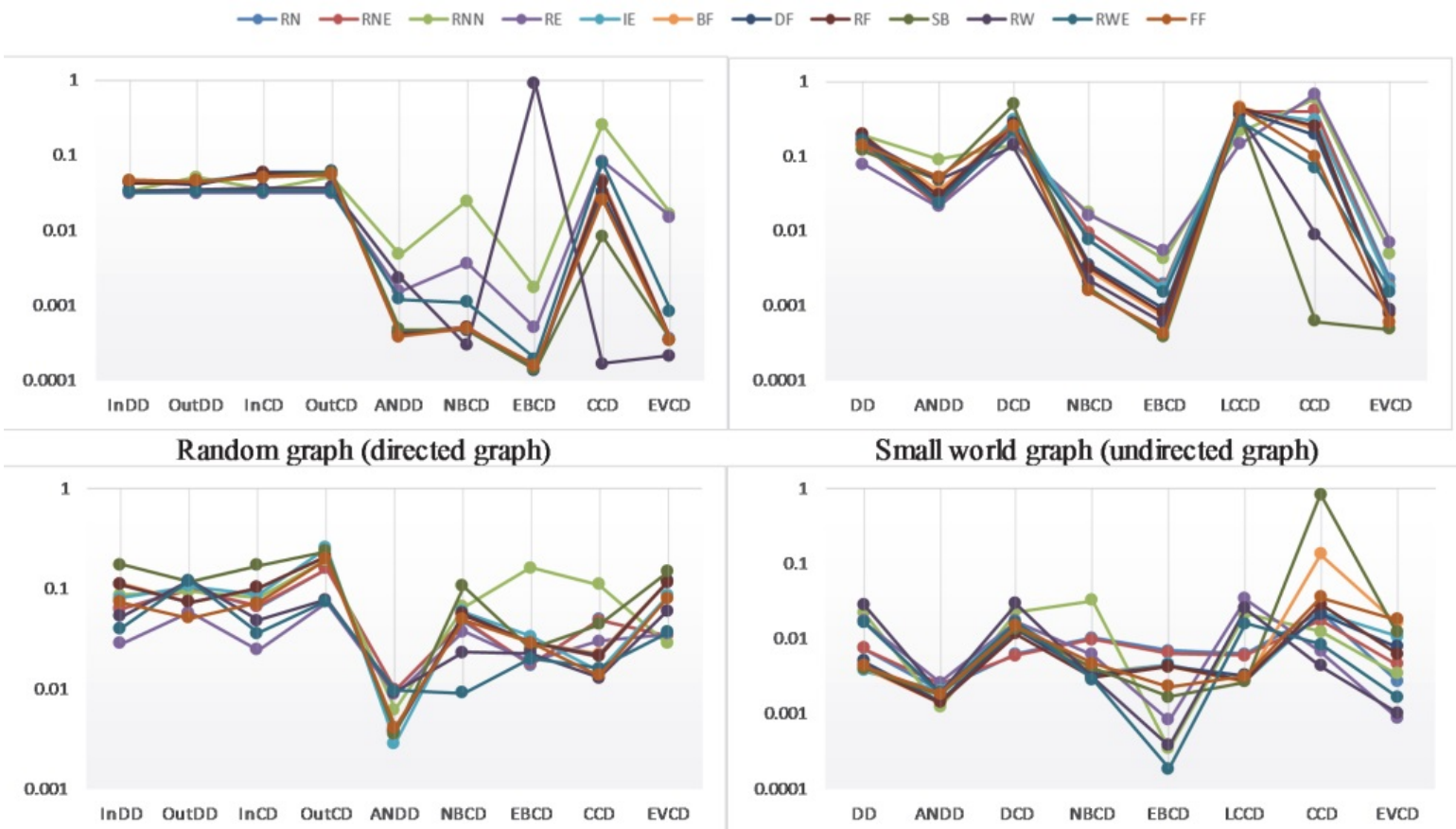

Figura: Resultado médio das comparações estatísticas entre os métodos de amostragem com taxas de amostragem de 10 a 50 por cento. O eixo vertical é os valores SD, o eixo horizontal representa as propriedades do gráfico e as linhas são métodos de amostragem.

Comparações estatísticas

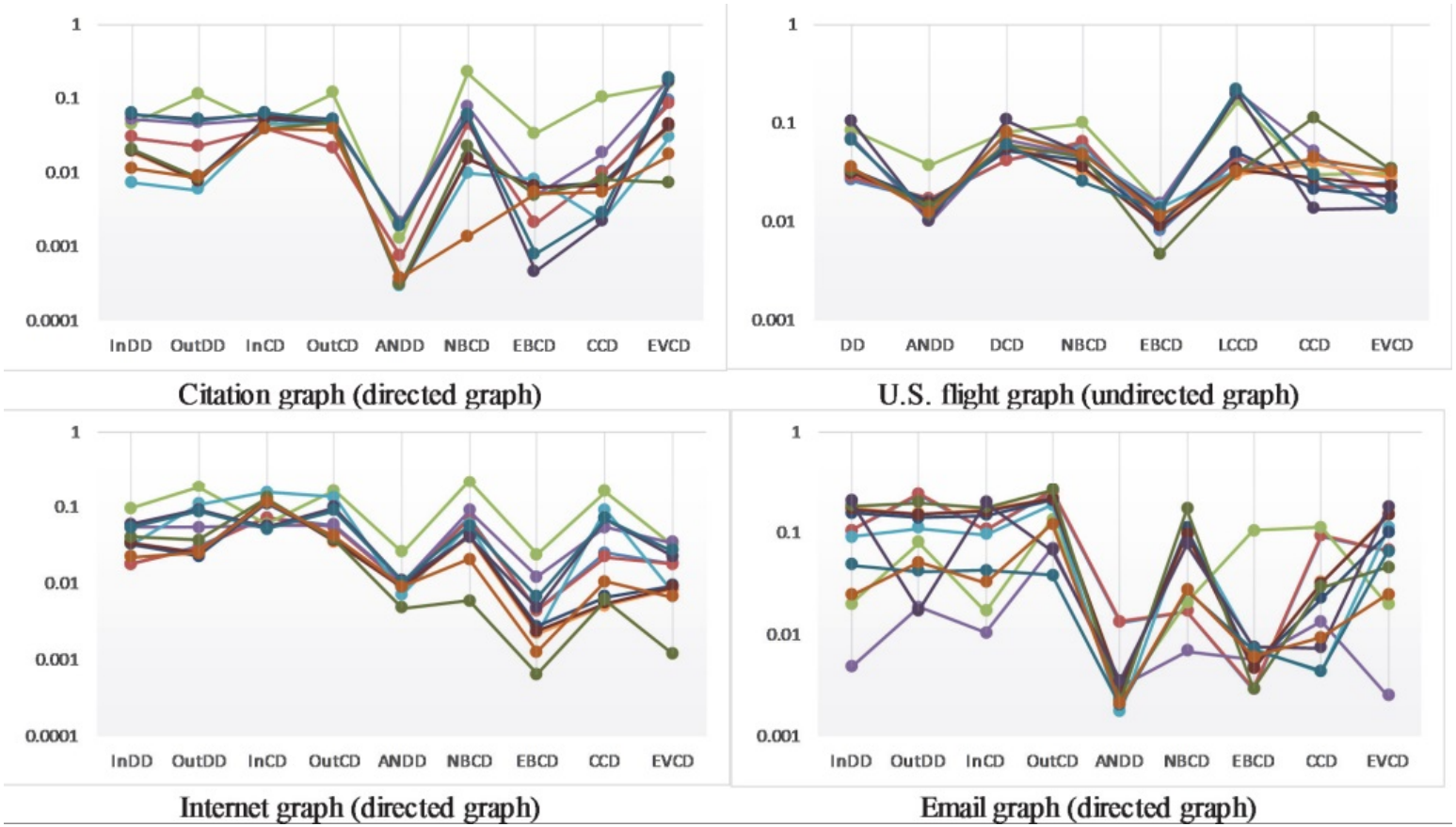

Figura: Resultado médio das comparações estatísticas entre os métodos de amostragem com taxas de amostragem de 10 a 50 por cento. O eixo vertical é os valores SD, o eixo horizontal representa as propriedades do gráfico e as linhas são métodos de amostragem.

Comparações estatísticas

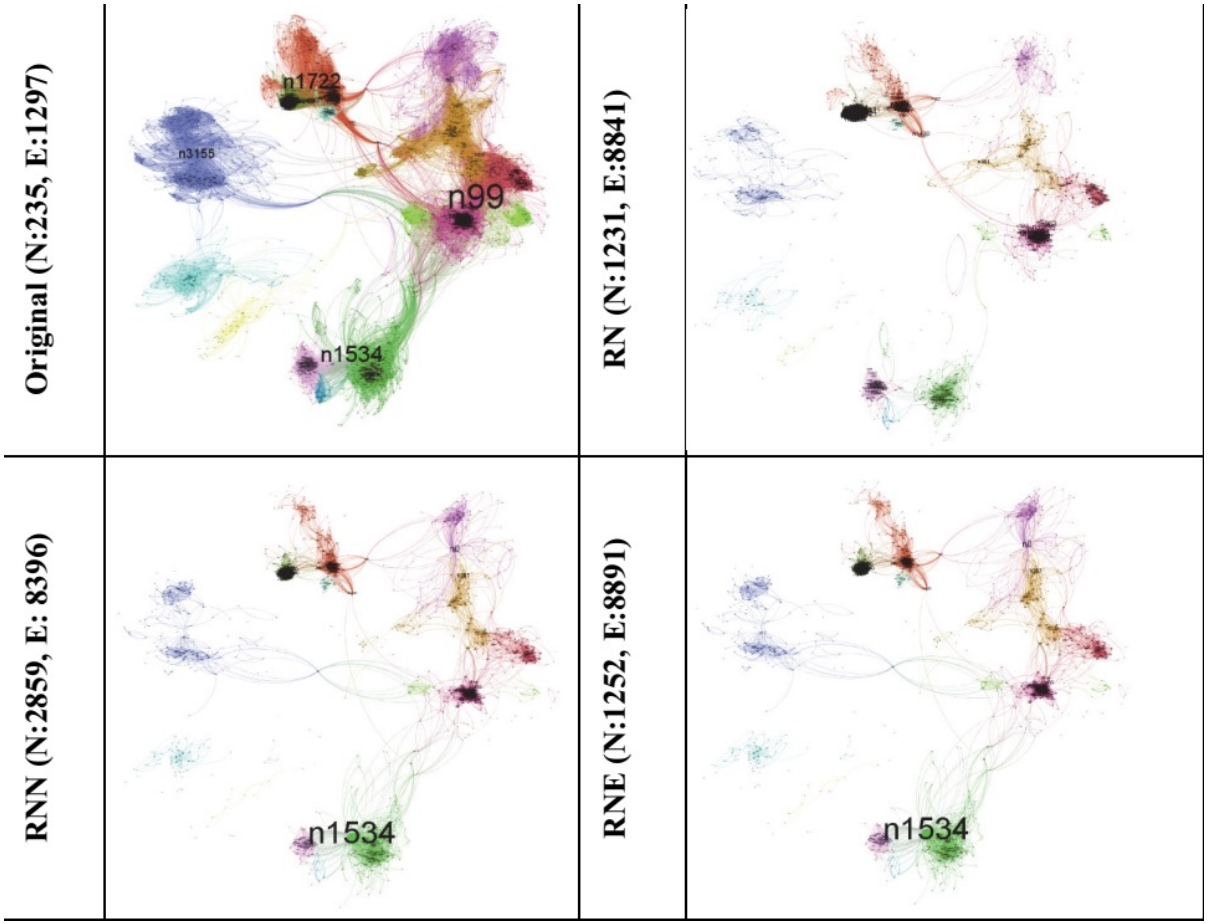

Figura:

Comparações estatísticas

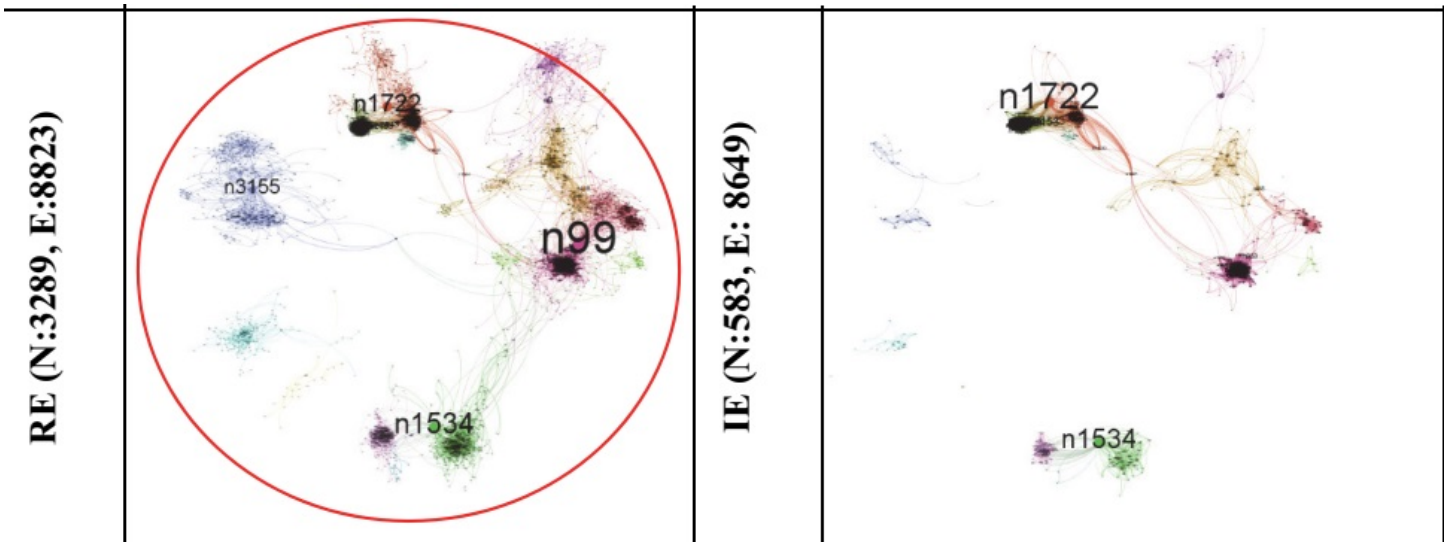

Figura: Comparação visual entre métodos de amostragem para dados do gráfico do Facebook (gráfico não direcionado) com taxa de amostragem de 10 porcentos nas bordas. Círculos Vermelhos no ER e círculo azul claro no SB mostram a área de cobertura espacial dos resultados da amostragem.

Comparações estatísticas

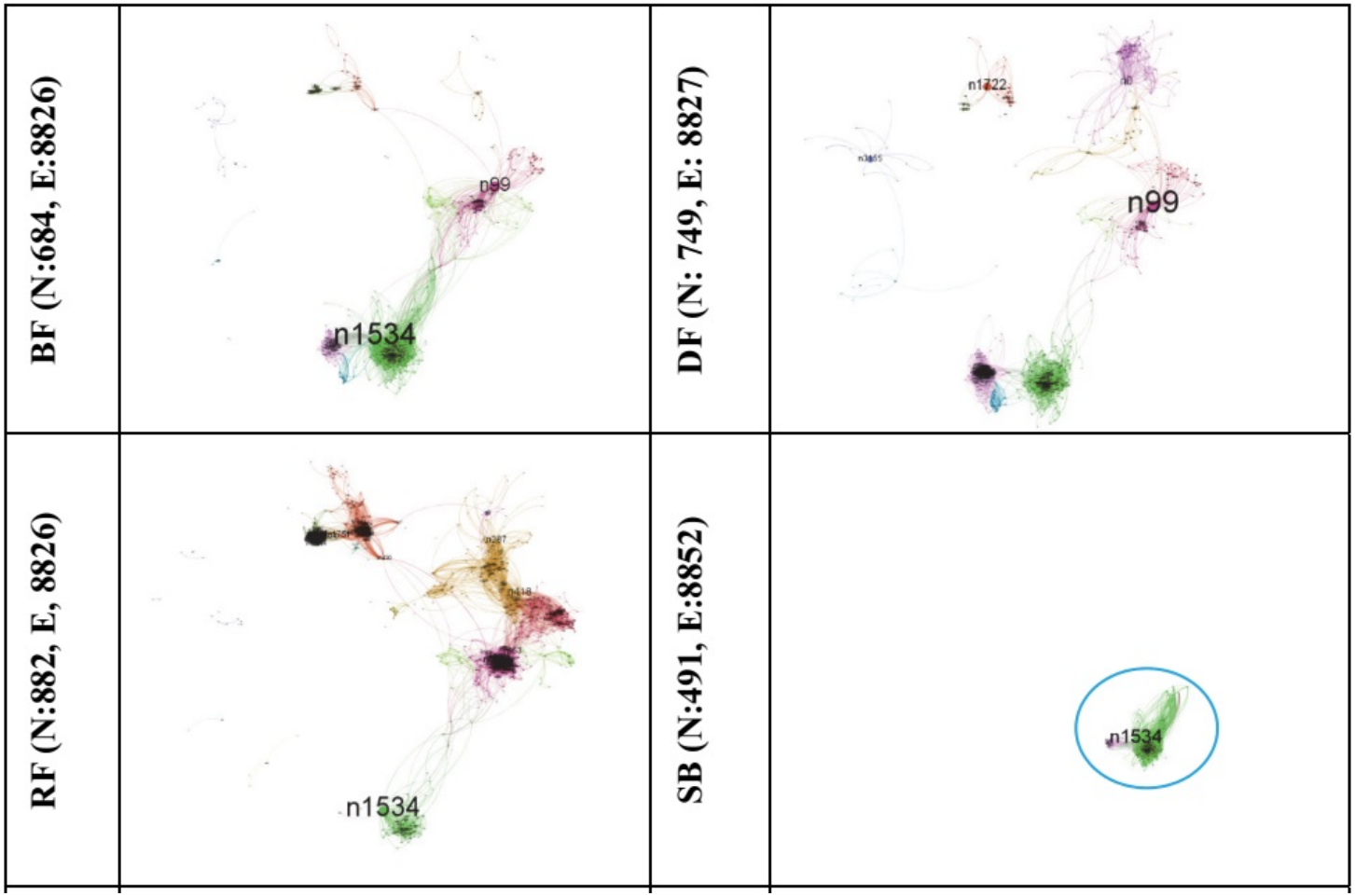

Figura:

Comparações estatísticas

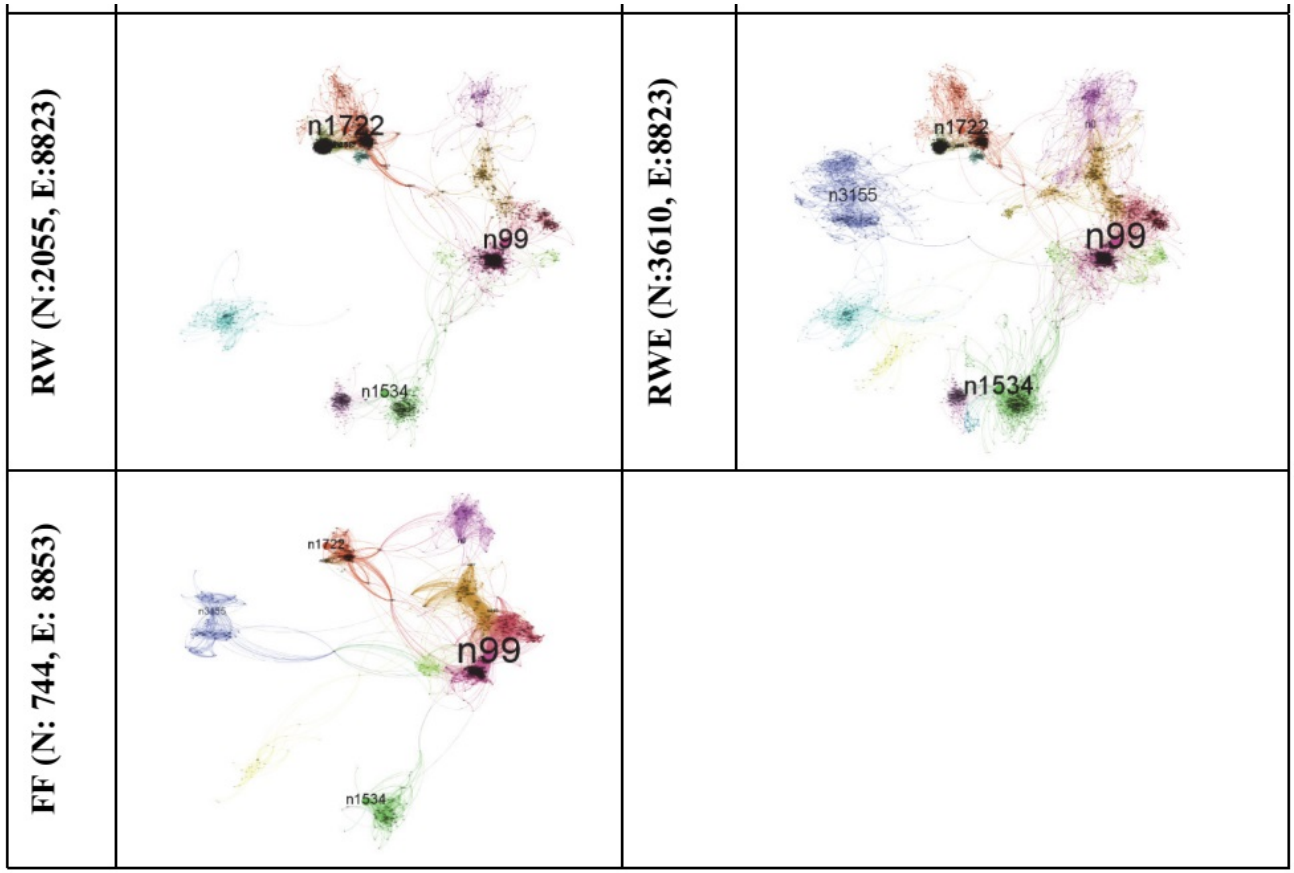

Figura: Comparação visual entre métodos de amostragem para dados do gráfico do Facebook (gráfico não direcionado) com taxa de amostragem de 10 porcentos nas bordas. Círculos Vermelhos no ER e círculo azul claro no SB mostram a área de cobertura espacial dos resultados da amostragem.

Comparações entre tipos de grafos

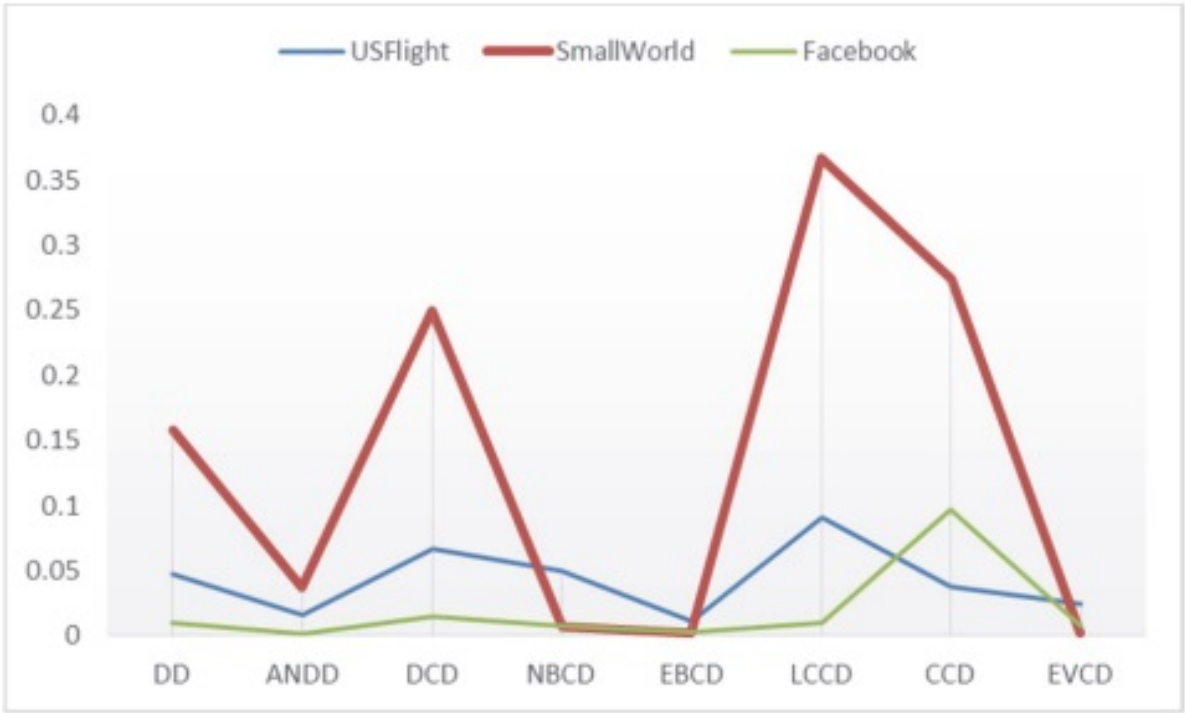

Figura: Resultados sumarizados para amostragem em grafos não direcionados.

Comparações entre tipos de grafos

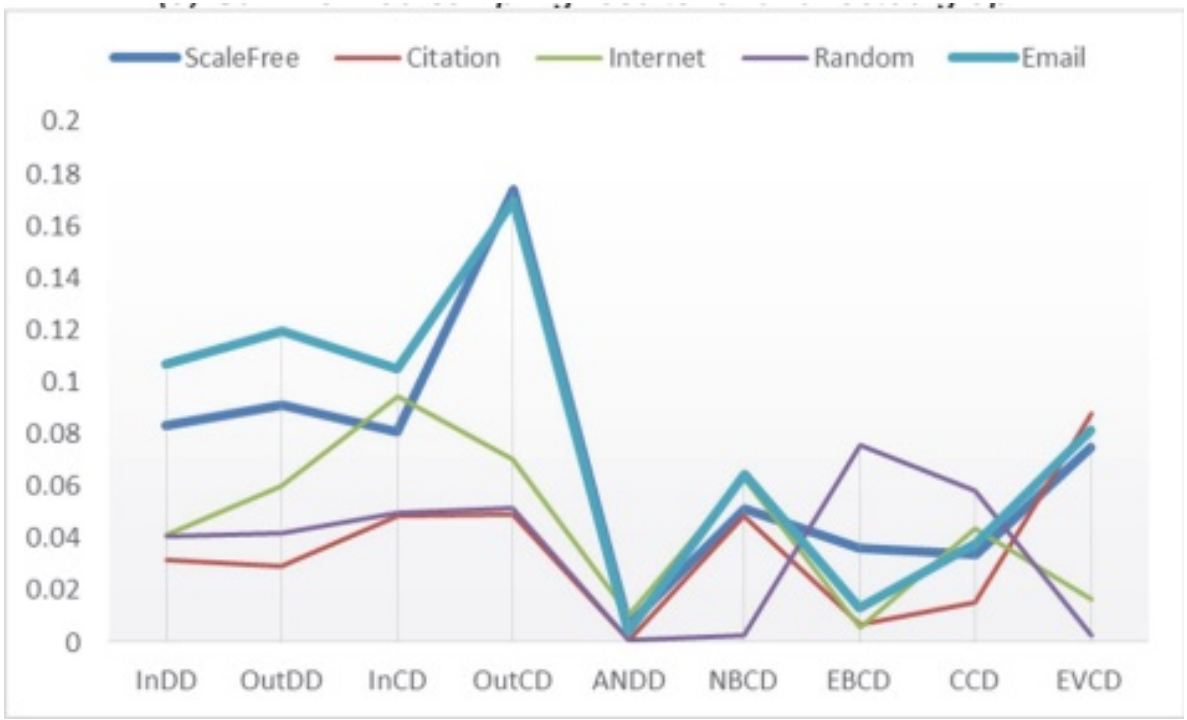

Figura: Resultados sumarizados para amostragem em grafos direcionados.

Comparações das propriedades

Apenas alguns métodos são consistentes para algumas propriedades de certos tipos de grafos.

- Amostragem Random walk conserva bem a **CCD** em grafos não direcionados;
- Amostragem induced-edge conserva a **ANDD** em grafos direcionados.

Alguns tipos de grafos preservam algumas propriedades muito bem. As redes email, scale-free e ramdom tem uma performance consistente para **DD, OutDD, InDD**.
Isto mostra que as propriedades do grafo devem ser consideradas na escolha da técnica de amostragem.

Comparações Visuais

Fornecem 2 critérios de comparação visual, e o quão boa é cada técnica de amostragem.

Critério de comparação visual

- spatial coverage/ cobertura espacial
  - Uma boa técnica de amostragem deve produzir um grafo com uma cobertura espacial similar a do grafo original

Dos resultados experimentais foi visto que métodos aleatórios (**random sampling**) tem melhor cobertura espacial que os métodos transversais (**traversal-based**). Como exemplo veja

na figura de comparações visuais O resultados de Ramdom sampling da rede Facebook, e traversal-based sampling tais como como snowball.

Comparações Visuais

- Critério do número de Clusters o tamanho,e a estrutura.
  - observa-se que os métodos de amostragem relacionados à aresta **(edge-related)** (por exemplo, random edge) são melhores que **node sampling** e **traversal-based** quando a taxa de amostragem é pequena

Observações Visuais

No artigo foram explorados 12 métodos de amostragem e aplicaram estes métodos a grafos de 235 a 265214 vértices e 1297 a 421578 arestas. Foram consideradas 9 propriedades para avaliar as técnicas de amostragem. A sua **referencia Visual e estatística** deve avaliar métodos de amostragem e a sua efetividade deve preservar as propriedades estatísticas e visuais do grafo original.

Observações Visuais

Os métodos de amostragem dependem de vários factores

- Tipo de grafo;
- Propriedade estatística desejada;
- Requisitos visuais

Os critérios de comparação visual fornecidos ajudam o usuário na comparação visual de grafos.

Sobre O Artigo

O Artigo explora as variadas técnicas de amostragem, estuda o impacto estatístico que cada um tem, do outro lado estuda o efeito visual que cada técnica produz, apresenta uma referencia de comparação visual e estatística mas não da uma explicação da razão da eficiência ou a falta de eficiência de uma determinada técnica de amostragem, tanto no sentido estatístico ou visual.

Algumas Questões

1. Diga duas razões para a amostragem sobre grafos;
2. Diga quais são os critérios de comparação Visual entre grafos;
3. Quais são os factores principais ao escolher o tipo de amostragem nos grafos.

Referencia do Artigo

- **Título**: A Visual Evaluation Study of Graph Sampling Techniques
- Volume: 63,
- data: 29 de Janeiro de 2017
- DOI: `https://doi.org/10.2352/ISSN.2470-1173.2017.1.VDA-394`
- Publisher: Society for Imaging Science and Technology
- Citações: 3 google Acadêmico

O artigo é da áreia de visualização, e as suas referencias são de da areia de visualização.

Agradecimento

Muito Obrigado!