Aula 2 - Introdução às ferramentas de análise

José Devezas <jld@fe.up.pt>
Faculdade de Engenharia da Universidade do Porto

Gestão da Informação em Redes Sociais
Mestrado em Ciência da Informação

Conteúdos

- Introdução
- Formatos de dados
- Conjuntos de dados
- Tutorial de Gephi
  - Visão geral
  - Carregar os dados
  - Organizar visualmente os nós
  - Calcular e visualizar estatísticas
  - Explorar e filtrar o grafo
  - Utilizar o “Data Laboratory”
  - Exportar visualizações

Introdução

- Ferramentas de análise de redes:
  - Gephi
  - Outros: Cytoscape, NodeXL, etc.
- Formatos de dados para representação de grafos:
  - Comma-Separated Value (CSV)
  - Graph Modeling Language (GML)
  - Outros: GEXF, GDF, GraphML, Pajek NET, etc.
- Caraterização de uma rede social com o Gephi.

Formatos de dados

- Comma-Separated Value (CSV)
  - Gephi: “File⇒Open”.
  - Texto simples, compatível com spreadsheets (p.e. Microsoft Excel).
  - **edge_list.csv**
    - source ⇒ id do nó de origem
    - target ⇒ id do nó de destino
  - Formato muito básico e que não suporta atributos (p.e. *label* para o tipo de aresta).

edge_list.csv

```
1,2
2,3
4,3
2,4
```

Formatos de dados

- Comma-Separated Value (CSV)
  - Gephi: “File⇒Import spreadsheet...”.
  - Texto simples, compatível com spreadsheets (p.e. Microsoft Excel).
  - **nodes.csv**
    - id ⇒ identificador único do nó
    - label ⇒ nome do nó
  - **edges.csv**
    - source ⇒ id do nó de origem
    - target ⇒ id do nó de destino
    - label ⇒ tipo de aresta
  - Formato mais flexível.

nodes.csv

```
id,label
1,"Arya Stark"
2,"John Snow"
3,"Daenerys Targaryen"
4,"Sansa Stark"
```

edges.csv

```
source,target,label
1,2,mentioned
2,3,retweeted
4,3,mentioned
2,4,retweeted
```

Formatos de dados

- Graph Modeling Language (GML)
  - Gephi: “File⇒Open”.
  - Formato de texto que combina nós e arestas.
  - node
    - id ⇒ identificador único do nó
    - label ⇒ nome do nó
  - edge
    - source ⇒ id do nó de origem
    - target ⇒ id do nó de destino
    - label ⇒ tipo de aresta
  - Suporta múltiplos atributos de nó e aresta (p.e. *text* com conteúdo textual associado a um nó)

graph.gml

```
graph [
  directed 1
  node
  [
    id 1
    label "Arya Stark"
  ]
  ...
  ...
  edge
  [
    source 1
    target 2
    label "mentioned"
  ]
  ...
]
```

Conjuntos de dados

- Existem diversas redes sociais disponíveis na web nos formatos descritos.
- Uma boa fonte para estes conjuntos de dados é o site do SNAP: http://snap.stanford.edu/data/
- Grafos com mais de 5.000 nós ou mais de 100.000 arestas são geralmente mais difíceis de tratar visualmente.
- Para a gestão da informação nas redes sociais, é necessário primeiro caraterizar e conhecer a rede.

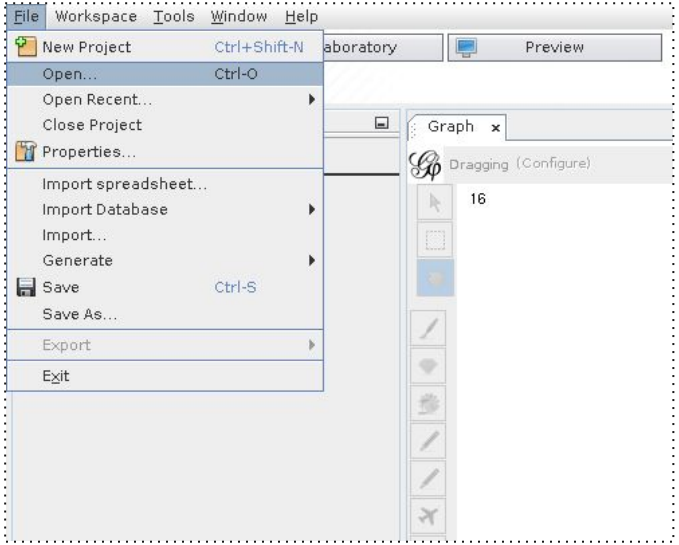
- Uma rede social é qualquer rede que se centra nas relações entre os indivíduos, sejam elas relativas a:
  - Amizades;
  - Seguidores;
  - Menções;
  - Partilhas;
  - Coautorias;
  - Ou outras interações.

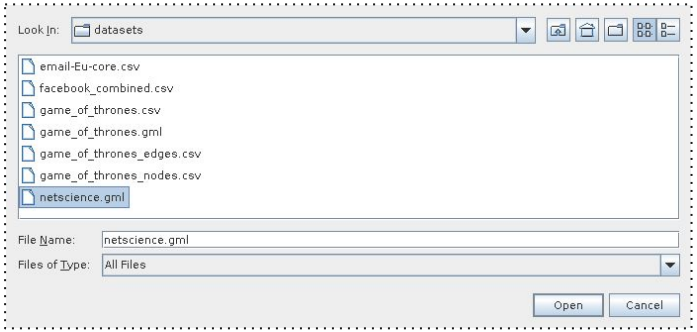
Tutorial de Gephi

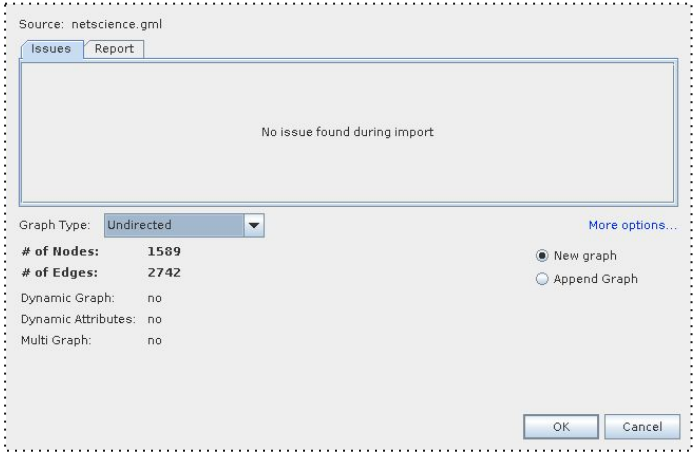
Visão geral

- Cada instância do Gephi mostra um projeto.
- Cada projeto pode ter vários “Workspaces”:
  - Com grafos diferentes;
  - Ou formatações diferentes do mesmo grafo.
- **ATENÇÃO!** Infelizmente, o Gephi não possui botão de “undo”.
  - A estratégia será fazer as formatações básicas do grafo, replicando-o em alguns “Workspaces”, caso seja necessário voltar atrás ou gerar uma nova versão a partir da versão-base.

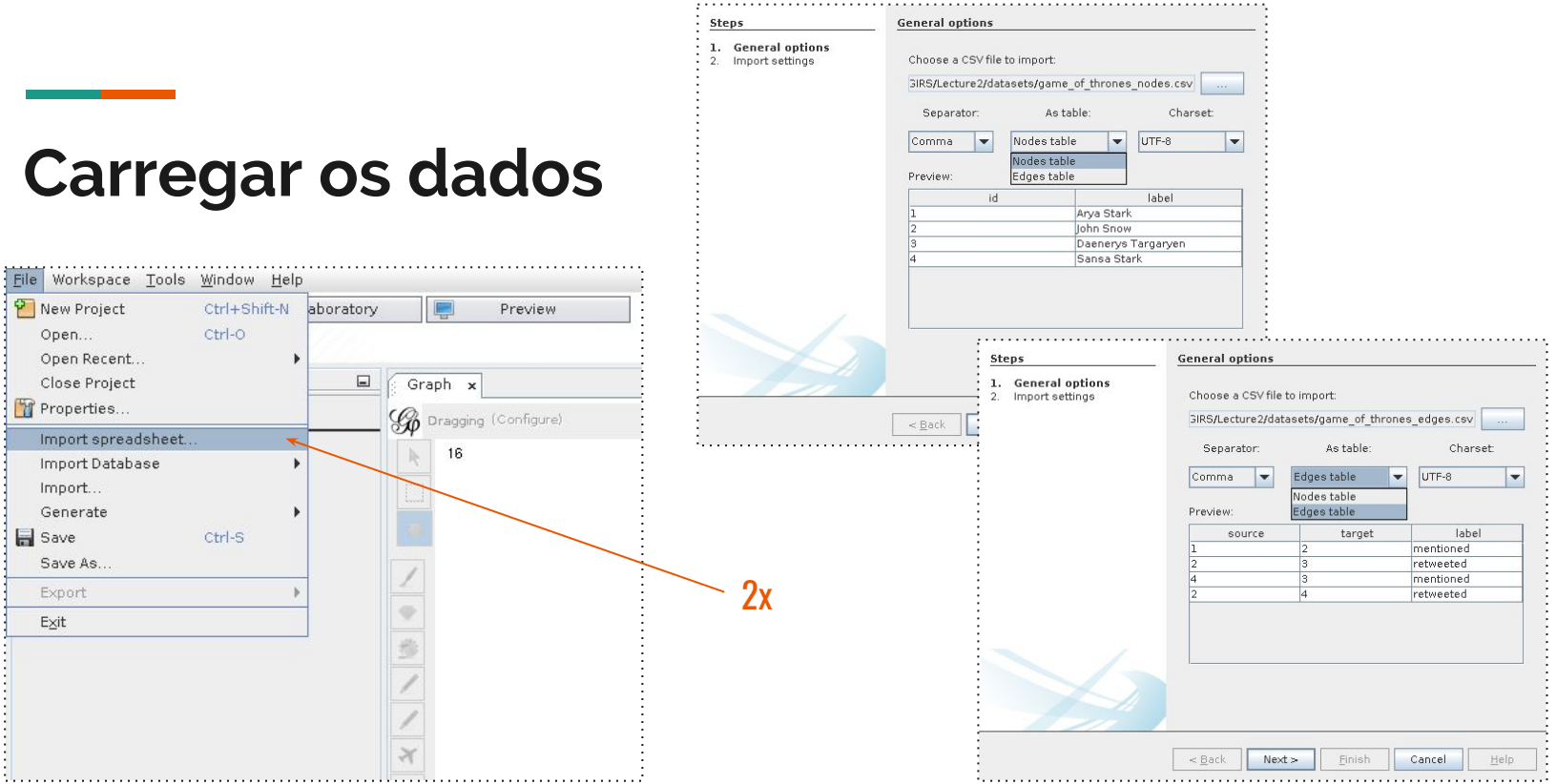
Carregar os dados





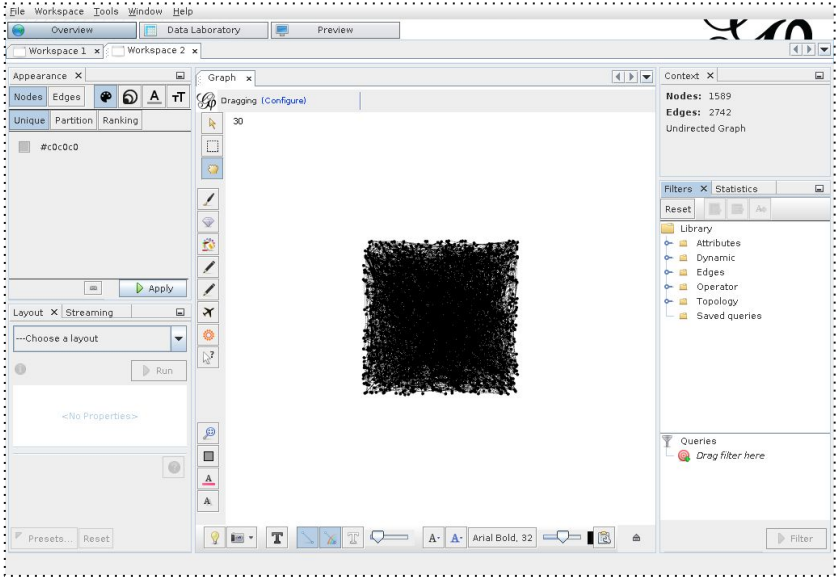


Carregar os dados



Organizar visualmente os nós

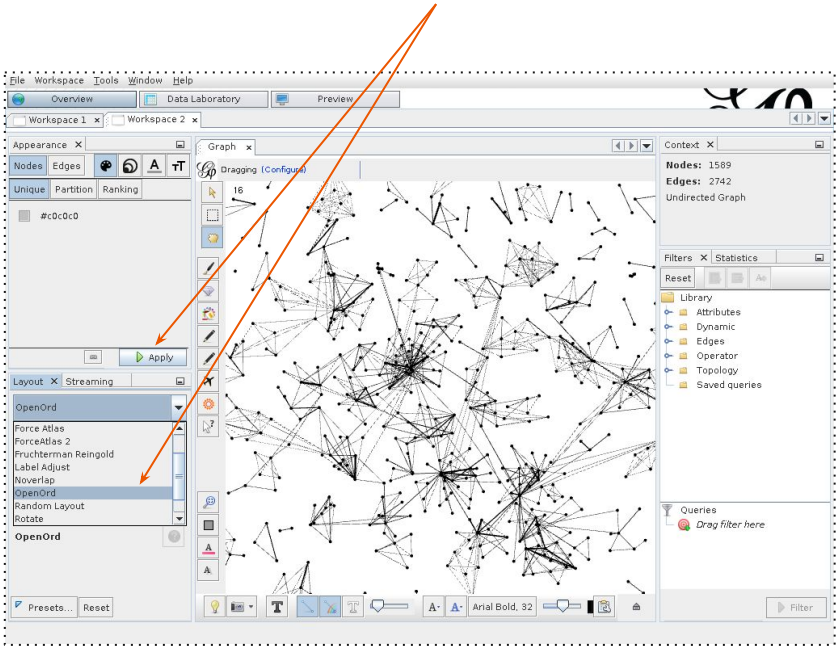
Ao carregar o grafo, obtemos uma representação que não oferece informação visual relevante.



Organizar visualmente os nós

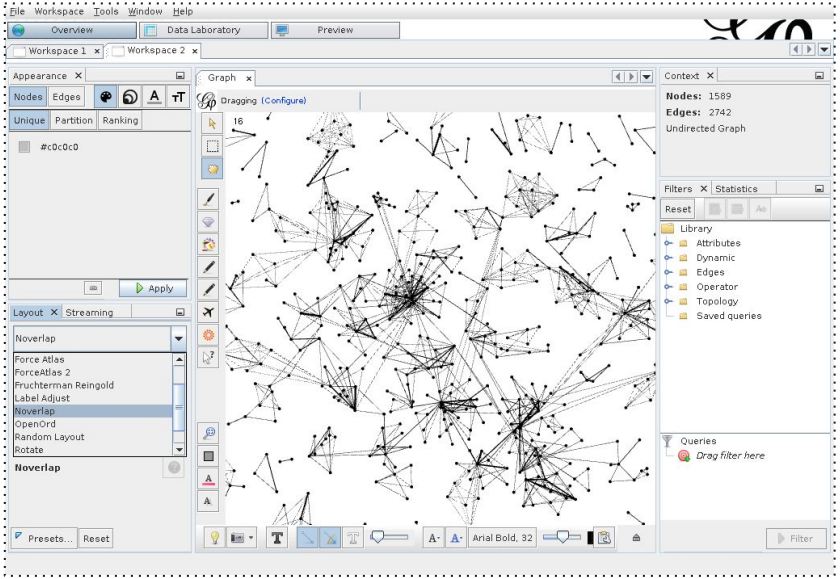
No separador “Layout”, encontramos várias opções para manipular automaticamente o posicionamento dos nós.

O “OpenOrd” é um algoritmo que permite reposicionar os nós do grafo, fazendo uma boa separação por módulos (comunidades).



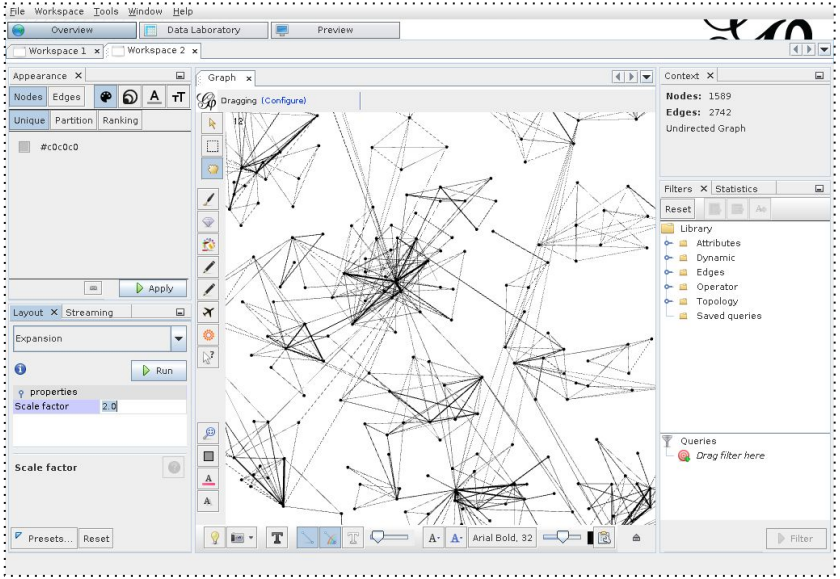
Organizar visualmente os nós

Como o “OpenOrd” tem tendência a sobrepor nós do mesmo módulo, podemos utilizar o “Noverlap” para evitar que alguns nós fiquem escondidos.



Organizar visualmente os nós

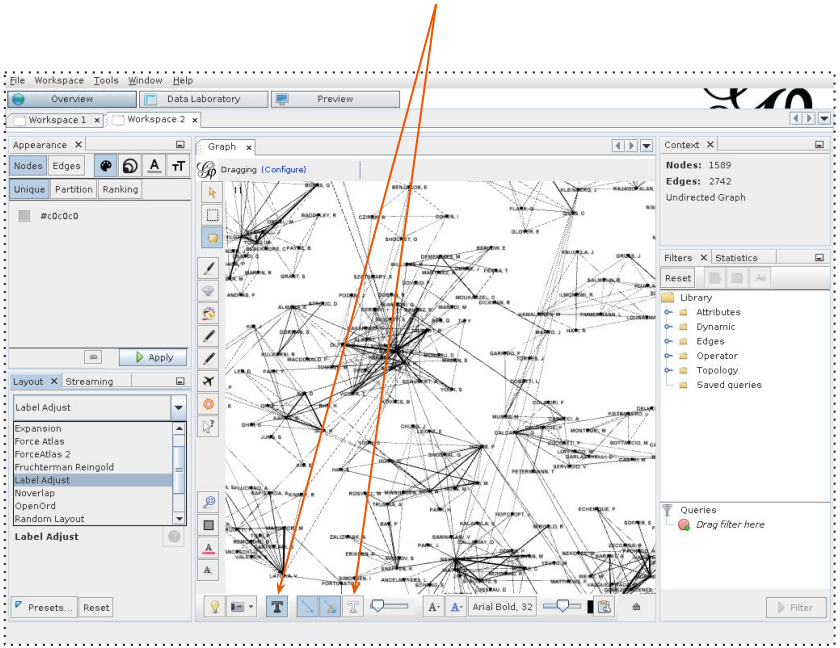
Vamos escalar um pouco a rede de forma a aumentar o espaço entre os nós, utilizando o “Expansion”, com “Scale factor” igual a 2.



Organizar visualmente os nós

No passo anterior, expandimos o grafo para podermos visualizar o nome dos nós de forma mais legível. Corremos o “Label Adjust” para evitar sobreposição do texto.

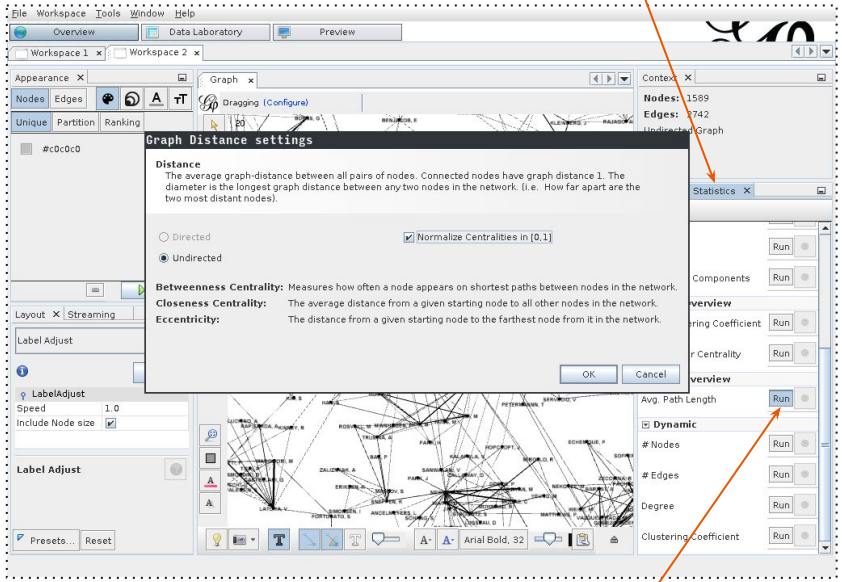
Os botões bom um ‘T’ servem para mostrar o atributo “Label” dos nós (a preto) e das arestas (a branco). Ligamos para os nós.



Calcular e visualizar estatísticas

Vamos calcular a *betweenness centrality*, normalizada entre 0 e 1. Se o grafo fosse dirigido, teríamos a opção de o tratar como não-dirigido.

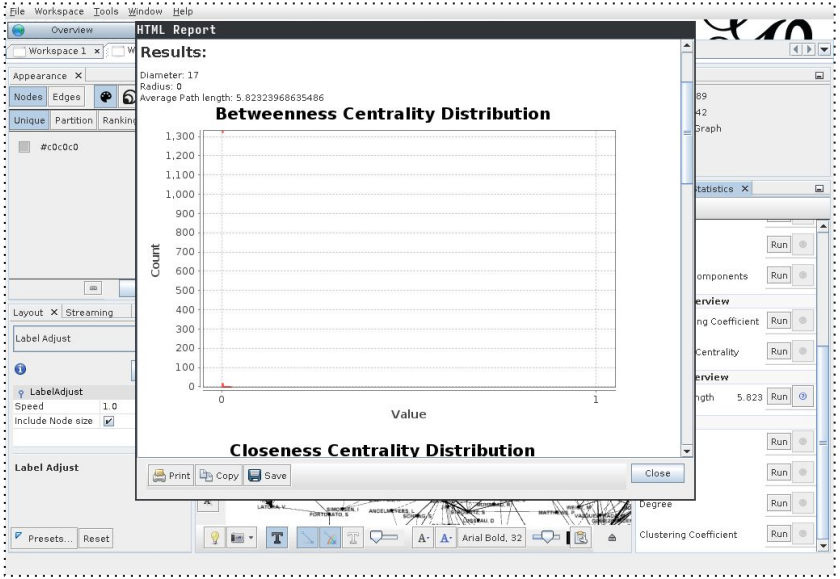
Selecionar separador “Statistics”.



Selecionar “Run” para “Avg. Path Length”.

Calcular e visualizar estatísticas

Os resultados mostram o diâmetro e o raio do grafo, o tamanho médio do caminho mais curto, e a distribuição das centralidades e da excentricidade.



Conceitos

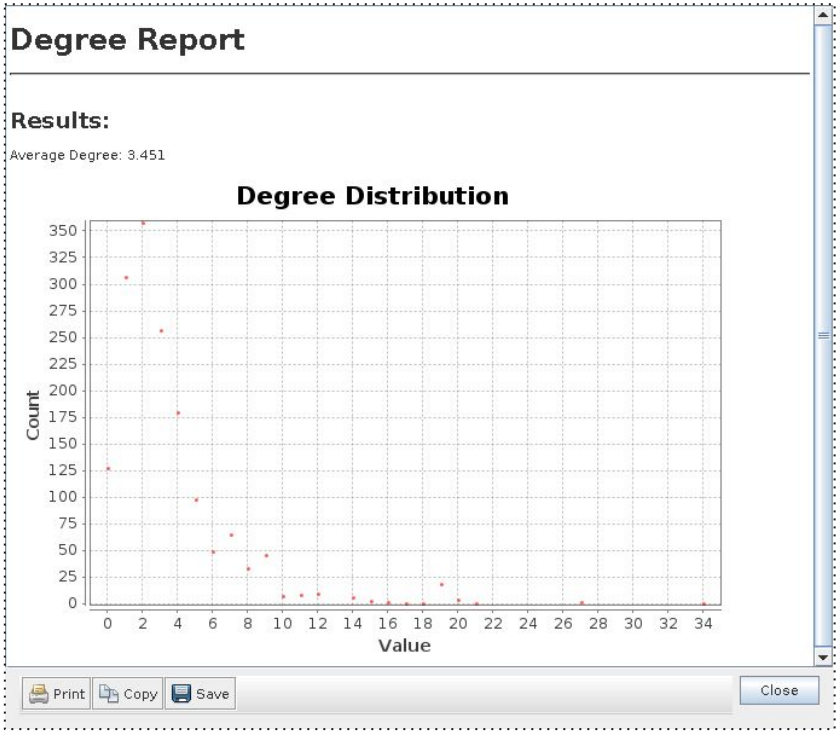
- **Distância geodésica ⇒** A distância geodésica entre dois nós é o número de arestas para o caminho mais curto que une esses dois nós.
- **Excentricidade ⇒** A excentricidade de um nó é igual à maior distância geodésica entre o nó e todos os outros nós.
- **Diâmetro ⇒** O diâmetro de um grafo corresponde ao valor da maior excentricidade.
- **Raio ⇒** O raio de um grafo corresponde ao valor da menor excentricidade.

Nota: O tamanho médio do caminho mais curto permite verificar o ***small-world effect***. São **seis graus de separação**? Ou o mundo ainda é mais pequeno, como no Facebook, em que são apenas quatro?

Calcular e visualizar estatísticas

Para termos também uma intuição sobre a ***scale-freedom*** da rede, isto é, se a distribuição do grau dos seus nós segue uma ***power law***, podemos consultar a "Degree Distribution", no relatório do "Average Degree", em "Statistics".

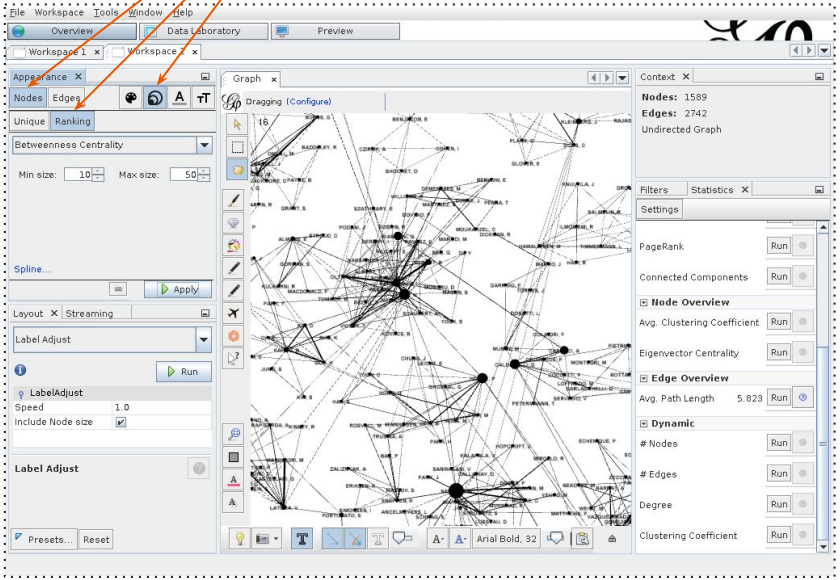
Dado que o gráfico não está na escala log-log, a função não é uma linha, mas tudo indica a que estamos na presença de uma *scale-free network*.



Calcular e visualizar estatísticas

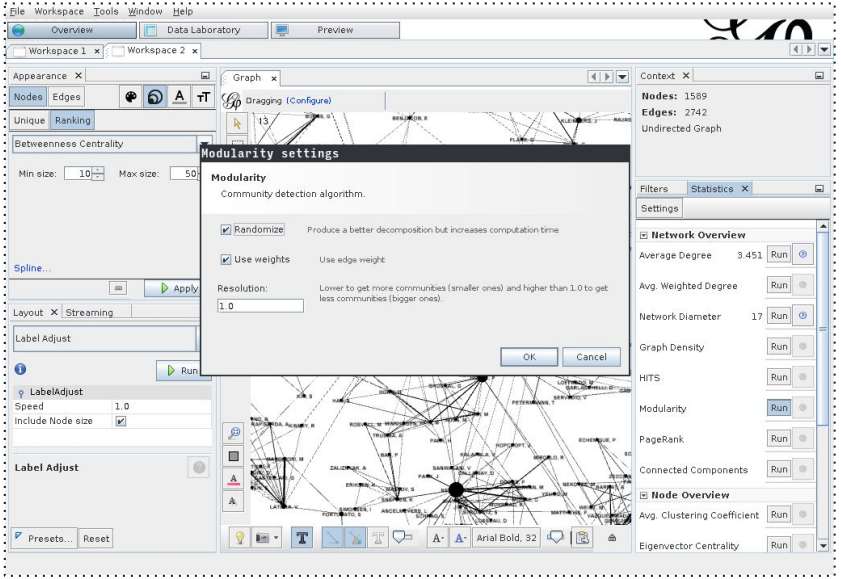
Vamos associar ao tamanho do nó a sua *betweenness centrality*, mapeando o menor valor para tamanho 10 e o maior valor para tamanho 50.

No separador “Appearance”, selecionar “Nodes”, seguido do ícone do tamanho e de “Ranking”.



Calcular e visualizar estatísticas

Agora vamos detectar comunidades com o método de Louvain, correndo a “Modularity”.

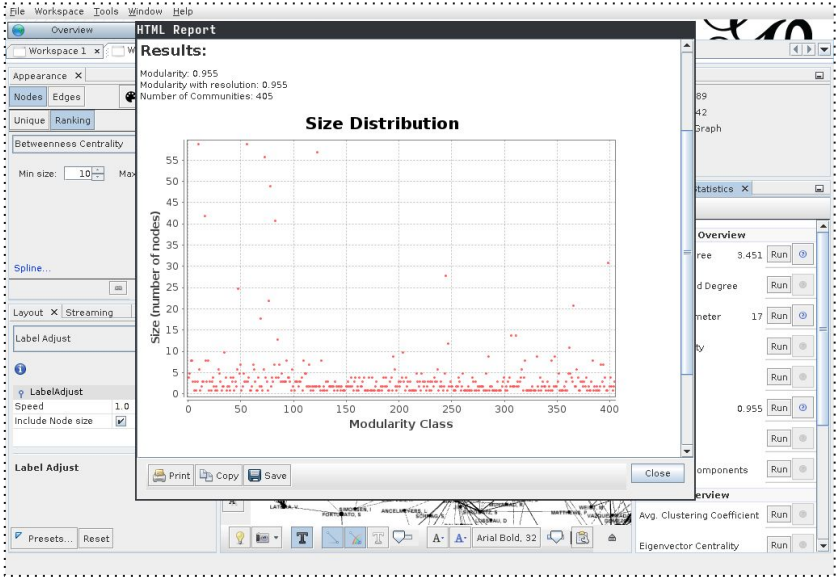


Conceitos

- **Modularidade ⇛** É um indicador da qualidade de uma partição do grafo, ou seja, mede a força de divisão da rede em módulos (comunidades).
  - É calculada através da fração de arestas que pertencem aos módulos menos a fração esperada de arestas, se estas tivessem sido distribuídas aleatoriamente.
  - Varia entre -½ e 1 (exclusive).
  - Numa rede real, é frequente a modularidade estar acima de 0.5 (não é uma regra, mas sim uma intuição).

Calcular e visualizar estatísticas

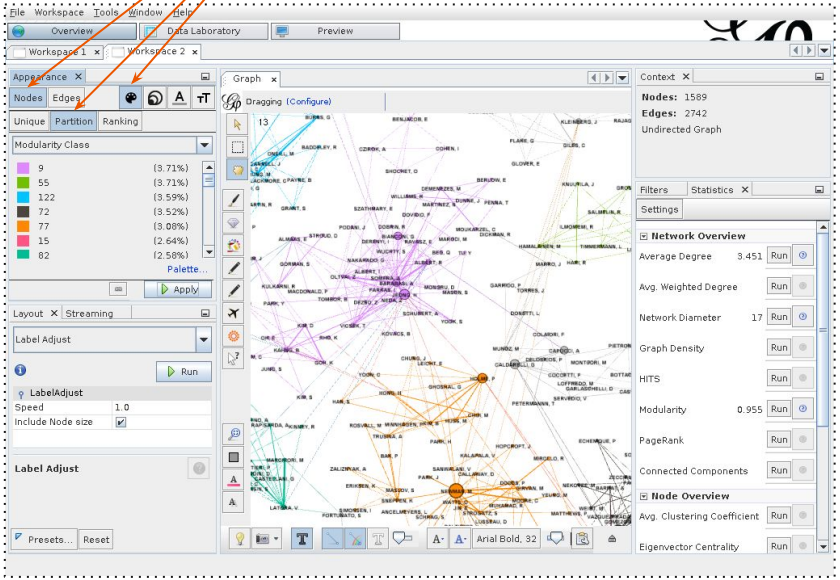
Os resultados mostram o valor da modularidade, o número de comunidades detectadas e a distribuição do tamanho (número de nós) das comunidades.



Calcular e visualizar estatísticas

Vamos associar à cor do nó a sua *comunidade*, mapeando cada classe para uma cor diferente. As comunidades mais pequenas são, por omissão, mapeadas para cinzento (+8 comunidades).

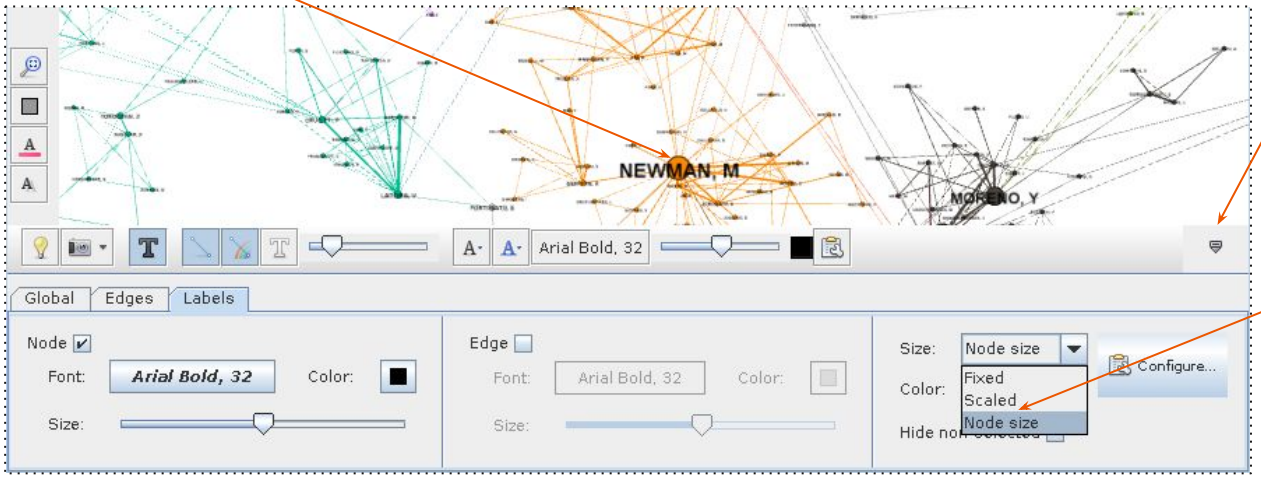
No separador “Appearance”, selecionar “Nodes”, seguido do ícone da cor e de “Partition”.



Calcular e visualizar estatísticas

Mark Newman aparece como um nó importante e sabemos que o é.

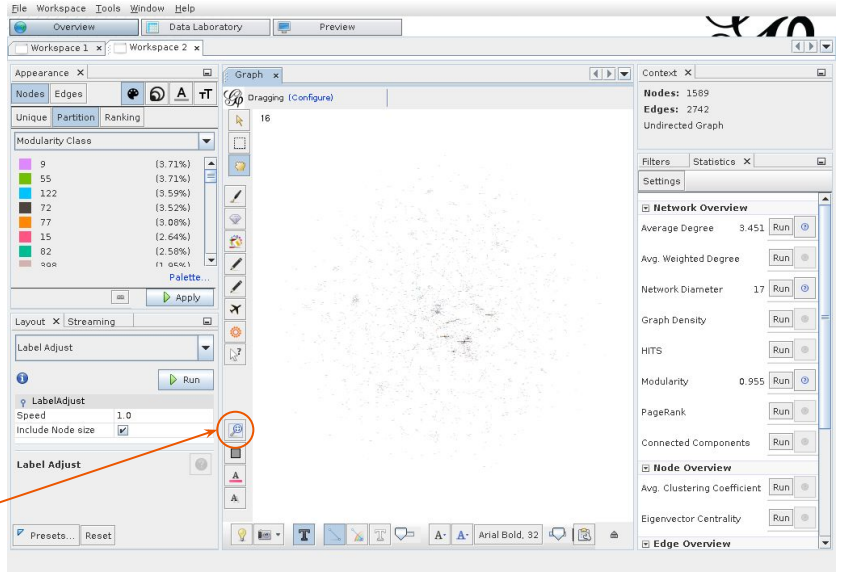
Mostrar ou esconder configurações de visualização.



No separador “Labels”, escolher o tamanho igual ao tamanho do nó, para realçar o nome dos nós mais importantes, segundo a *betweenness centrality*.

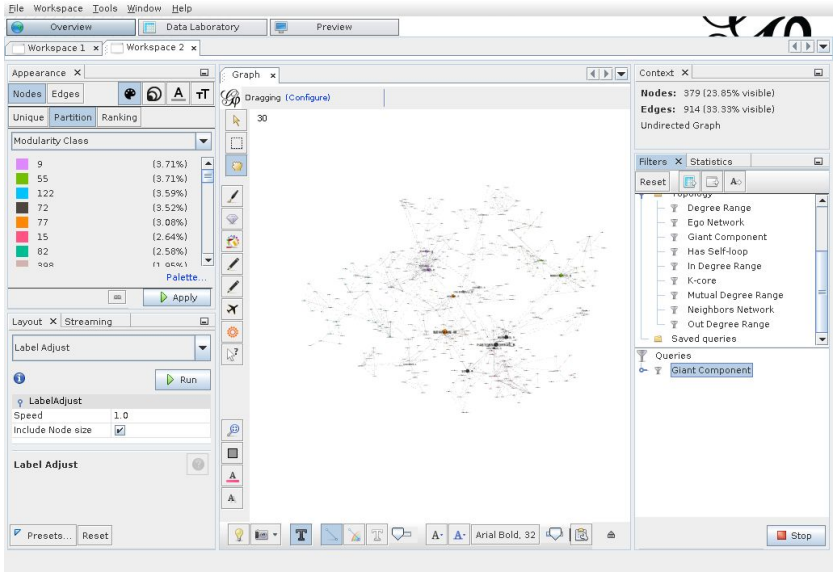
Filtrar e explorar o grafo

Ao explorar o grafo é possível usar o scroll para fazer zoom e o botão direito para arrastar a tela. Se o perdermos de vista, podemos clicar no botão “Center On Graph”.



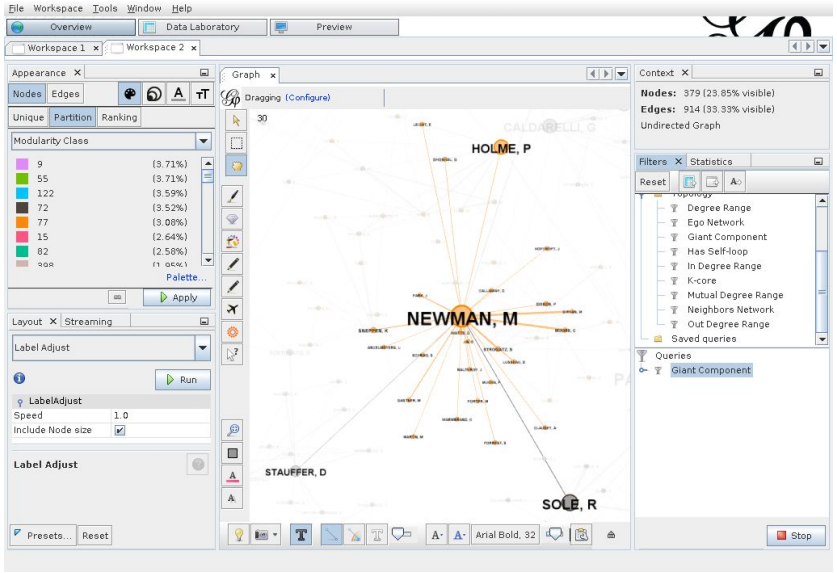
Filtrar e explorar o grafo

Para focarmos a atenção, podemos aplicar uma série de filtros, por exemplo o “Giant Component”, para ver apenas o maior componente ligado.



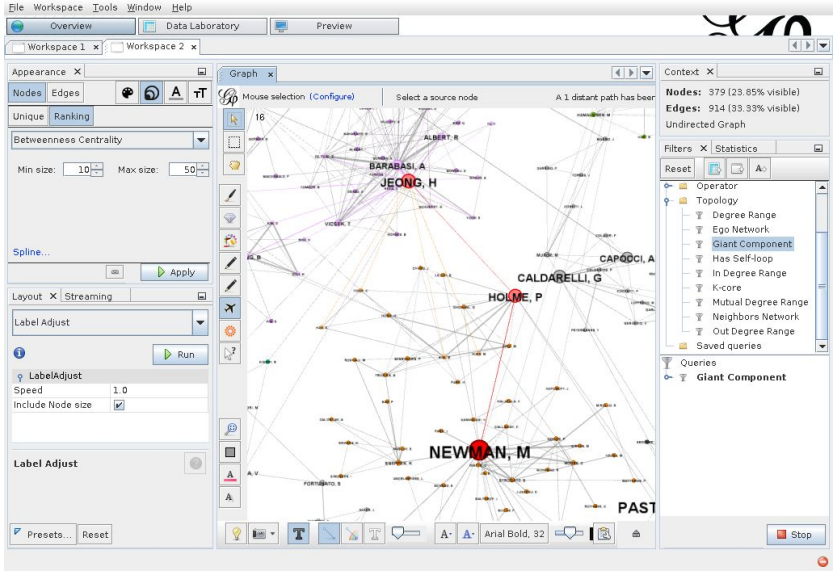
Filtrar e explorar o grafo

Fazer zoom e colocar o rato em cima no nó do Newman, para visualizar a sua vizinhança.



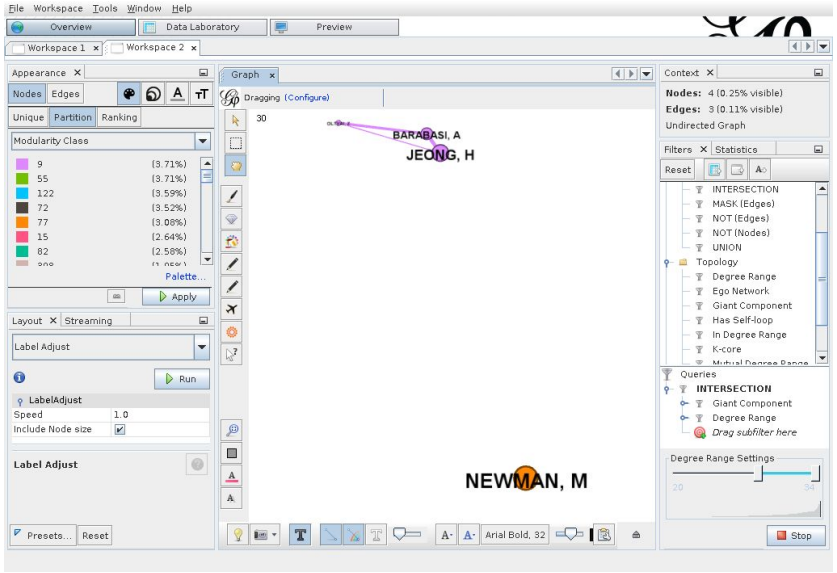
Filtrar e explorar o grafo

Utilizar o ícone em forma de avião para calcular o caminho mais curto entre dois nó, pintando todos os nós pertencentes ao caminho (a vermelho, na figura).



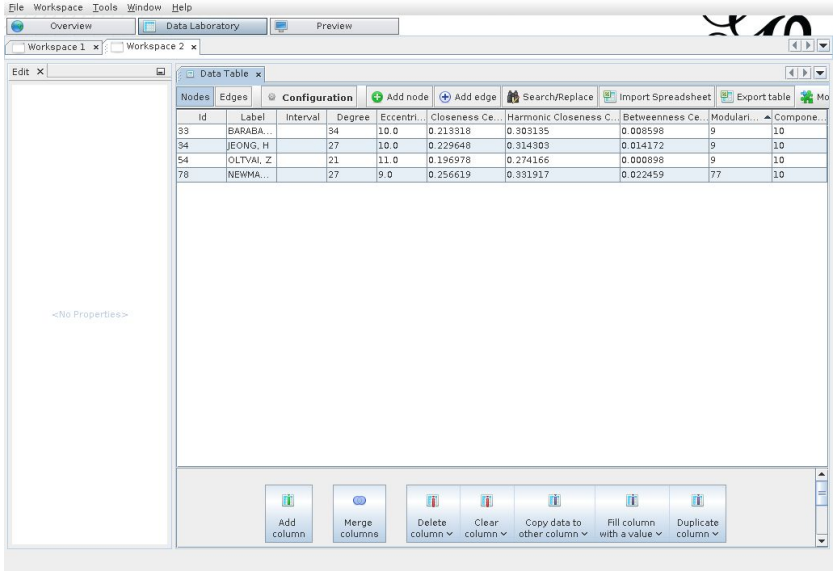
Filtrar e explorar o grafo

Podemos utilizar filtros compostos, por exemplo utilizando “INTERSECTION” com o “Giant Component” e o “Degree Range”, de forma a mostrar os nós do maior componente ligado com grau superior a 20 (no grafo original).



Utilizar o “Data Laboratory”

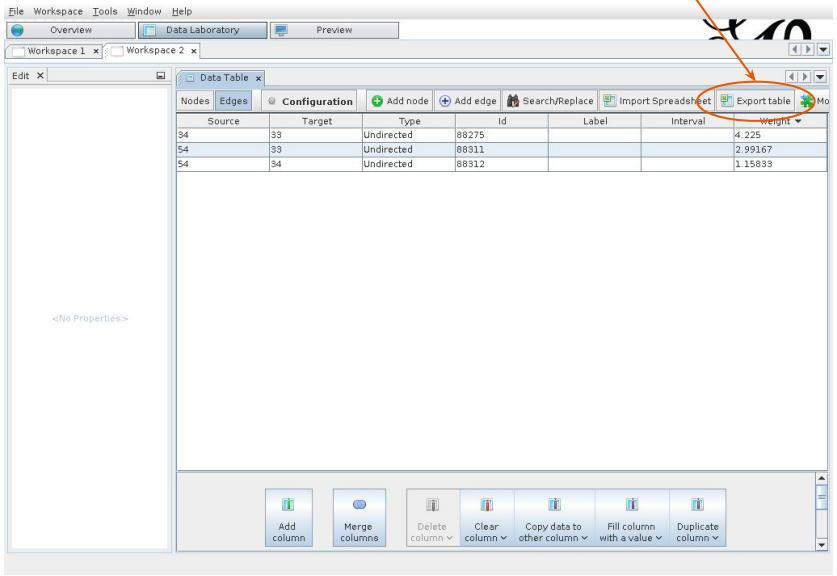
Permite explorar informação para os nós visíveis no “Overview” (os filtros têm impacto aqui). Na figura, vemos a informação para os nós, ordenada pela “Modularity Class” (o identificador da comunidade).



Utilizar o “Data Laboratory”

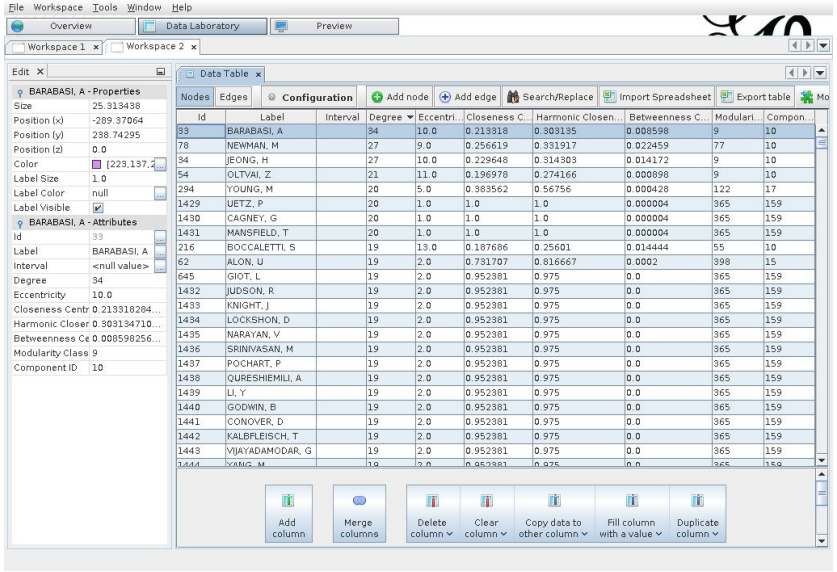
Aqui vemos o mesmo para as arestas, mostrando a informação ordenada pelo peso da aresta. Uma funcionalidade útil do “Data Laboratory” é a exportação para spreadsheet.

Exportar para spreadsheet em formato CSV.



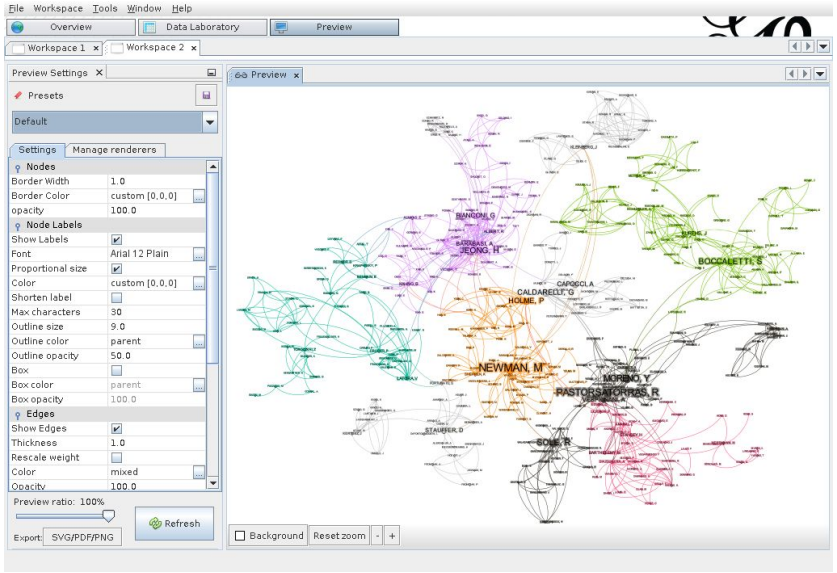
Utilizar o “Data Laboratory”

Ordenando por cada atributo de centralidade, conseguimos saber que: (1) o Albert-László Barabási tem o maior número de interações, (2) existem vários nós com com proximidade máxima, ou (3) o Mark Newman faz a ponte entre vários outros autores.



Exportar visualizações

Em “Preview”, é possível configurar a visualização. Em “Node Labels”, ligamos “Show Labels”, utilizamos 9.0 para “Outline size”, “parent” em “Outline color” e 50.0 em “Outline opacity”.



Exportar visualizações

É possível exportar para PDF, SVG ou PNG. Mostramos o resultado final, em PNG, na figura ao lado, apenas para uma pequena área do grafo (cortado num editor de imagem).