

# Aula 2 - Introdução às ferramentas de análise

José Devezas <<u>jld@fe.up.pt</u>> 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:
  - o 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  $\Rightarrow$  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  $\Rightarrow$  id do nó de origem
    - target  $\Rightarrow$  id do nó de destino
    - label  $\Rightarrow$  tipo de aresta
  - Formato mais flexível.

2		••
	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  $\Rightarrow$  identificador único do nó
    - label ⇒ nome do nó
  - edge
    - source ⇒ id do nó de origem
    - target  $\Rightarrow$  id do nó de destino
    - label  $\Rightarrow$  tipo de aresta
  - Suporta múltiplos atributos de nó e aresta (p.e. *text* com conteúdo textual associado a um nó)



#### 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: <u>http://snap.stanford.edu/data/</u>
- 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**

0	)pen )pen Recent	Ctrl-0	144	
o c	pen Recent	•		
P P	lose Project roperties			Graph ×
Ir Ir G S	nport spreadsheet nport Database nport Senerate Save	Ctrl-S		16
E	xport	•	- 1	/
E	⊻it		.9	2

	sets			00
nemail-Eu-core.	csv			
hacebook com	bined.csv			
 ngame of thror	ies.csv			
game of thron	ies.gml			
game of thron	ies edges.csv			
game of thron	ies nodes.csv			
netscience.gm	1			
	-			
ile <u>N</u> ame: net	science.gml			
iles of <u>T</u> ype: All	Files			
11				
			Open	Cance
Source: netscien	ort			
Source: netscien	re.gmi	No issue found du	ring import	
Source: netscient Tesues Repo Graph Type: Ur	rdirected	No issue found du	ring import More o	ption
Source: netscien Issues Repo Sraph Type: Ur # of Nodes:	idirected 1589	No issue found du	ring import More o @ New graph	ption
Source: netscien Tissues Repo Graph Type: Ur # of Nodes: # of Edges:	directed 1589 2742	No issue found du	ring import More o @ New graph _ Append Grap	ption
Source: netscien Tssues Repo Graph Type: Ur # of Nodes: # of Edges: Dynamic Graph:	ndirected 1589 2742 no	No issue found du	ring import More o @ New graph \_ Append Grap	ption
Source: netscien Issues Repo Graph Type: Ur # of Nodes: # of Edges: Dynamic Graph: Dynamic Attribut	directed 1589 2742 no es: no	No issue found du	ring import More o @ New graph _ Append Grap	ption
Source: netscien Issues Repo Sraph Type: Ur # of Nodes: # of Edges: Dynamic Graph: Dynamic Attributu Multi Graph:	directed 1589 2742 no sc: no no no	No issue found du	ring import More o @ New graph \_ Append Grap	ption
Source: netscien Tissues Repo Graph Type: Ur # of Nodes: # of Edges: Dynamic Graph: Dynamic Attribute Multi Graph:	directed 1589 2742 no es: no no	No issue found du	ring import More o @ New graph @ Append Grap	ption



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



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).



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



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



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.



#### Selecionar separador "Statistics".

# 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 "Run" para "Avg. Path Length".

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**  $\Rightarrow$  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?

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*.



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



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.
  - $\circ$  Varia entre - $\frac{1}{2}$  e 1 (exclusive).
  - Numa rede real, é frequente a modularidade estar acima de 0.5 (não é uma regra, mas sim uma intuição).

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



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).



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*.

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".



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



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



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).



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).

Exportable Acompone I0 I0 I0 I0 I0 I0 I0	Mi
dular Compone. 10 10 10 10 10 10	e
10 10 10 10	
10 10 10	
10 10 10	
10	
	-
	=
	H
1	



### 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.

Overview [	Data Laboratory	Preview					7	K //
Workspace 1 × Worksp	ace 2 ×	a						ब
lit ×	Data Table 🗴							1
	Nodes Edges	Configuration	🔂 Add node 🤙	Add edge 👪 Sea	arch/Replace	Import Spre	adsheet 関	Export table
	Source	Target	Туре	Id	Label	1	nterval	Weight
	34	33	Undirected	88275				4.225
	54	33	Undirected	88311				2.99167
	54	34	Undirected	88312				1.15833
<no properties=""></no>								
<no properties=""></no>								
<no properties=""></no>				7	16	li		
<no properties=""></no>		ii o		п	ĩ	li	ili,	

# 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.

Overvi	ew [ [	Data Labora	atory	Preview						X		
Workspace	1 × Workspac	e2 x									•	
Edit ×		Date	a Table 🗙								4	-
P BARABASI,	A - Properties	Nodes	Edges @ Conf	quiration	G Add n	ode 🕀 A	dd edge 🛛 🗰	Search/Replace	Import Spread	lsheet 🖳 Exp	ort table	E M
Size	25.313438			garation								
Position (x)	-289.37064	ld	Label	Interval	Degree •	Eccentri.	. Closeness	C Harmonic Clo	sen Betweenne	ss C Modulari	Compon.	
Position (y)	238.74295	33	BARABASI, A		34	10.0	0.213318	0.303135	0.008598	9	10	
Position (z)	0.0	78	NEWMAN, M		27	9.0	0.256619	0.331917	0.022459	11	10	-
Color	[223,137,2]	34	JEONG, H	-	27	10.0	0.229648	0.314303	0.014172	9	10	
Label Size	1.0	54	OLTVAI, Z		21	11.0	0.196978	0.274166	0.000898	9	10	
Label Color	null 🛄	294	YOUNG, M		20	5.0	0.383562	0.56756	0.000428	122	17	
Label Visible	2	1429	UETZ, P		20	1.0	1.0	1.0	0.000004	365	159	
P BARABASI,	A - Attributes	1430	CAGNEY, G		20	1.0	1.0	1.0	0.000004	365	159	
Id	33	1431	MANSFIELD, T		20	1.0	1.0	1.0	0.000004	365	159	
Label	BARABASI, A	216	BOCCALETTI, S		19	13.0	0.187686	0.25601	0.014444	55	10	
Interval	<null value=""></null>	62	ALON, U		19	2.0	0.731707	0.816667	0.0002	398	15	
Degree	34	645	GIOT, L		19	2.0	0.952381	0.975	0.0	365	159	
Eccentricity	10.0	1432	JUDSON, R		19	2.0	0.952381	0.975	0.0	365	159	Τ.
Closeness Cer	ntr 0 213318284	1433	KNIGHT, J		19	2.0	0.952381	0.975	0.0	365	159	
Harmonic Clos	er 0 303134710	1434	LOCKSHON, D		19	2.0	0.952381	0.975	0.0	365	159	
Retweenness	Ce 0.008598256	1435	NARAYAN, V		19	2.0	0.952381	0.975	0.0	365	159	
Modularity Cla	00 0.0000000200	1436	SRINIVASAN, M		19	2.0	0.952381	0.975	0.0	365	159	
Component ID	10	1437	POCHART, P		19	2.0	0.952381	0.975	0.0	365	159	
componentie		1438	QURESHIEMILI, A		19	2.0	0.952381	0.975	0.0	365	159	
		1439	LI, Y		19	2.0	0.952381	0.975	0.0	365	159	
		1440	GODWIN, B		19	2.0	0.952381	0.975	0.0	365	159	
		1441	CONOVER, D		19	2.0	0.952381	0.975	0.0	365	159	
		1442	KALBFLEISCH, T		19	2.0	0.952381	0.975	0.0	365	159	
		1443	VIIAYADAMODAR.	3	19	2.0	0.952381	0.975	0.0	365	159	
		1444	YANG M		19	2.0	0.952981	0.975	0.0	9.65	15.9	
												-
			<b>II</b>					<u>ni</u>	<u>II</u>	<b>II</b>		-
			Add	Mer	ge	Delete	Clear	Copy data to	Fill column D	uplicate		

# 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).

