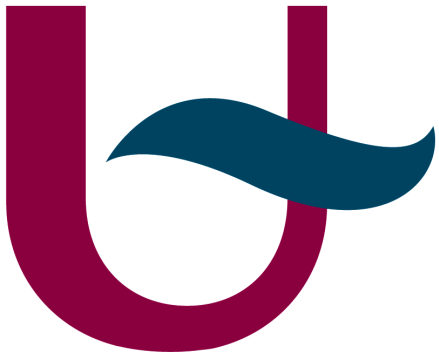
Unevenness in network properties

on the social Semantic Web



Universiteit Antwerpen

Raf Guns

Informatie- en Bibliotheekwetenschap

2nd Workshop on Social Aspects of the Web, 2008

Overview

Introduction: the Semantic Web

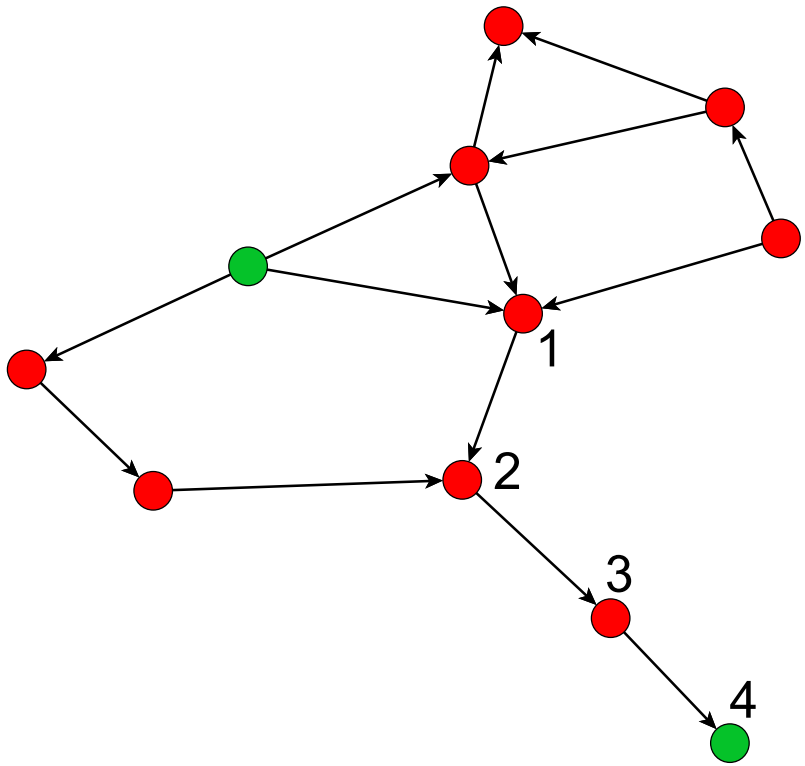
- *Social Semantic Web*: RDF data containing social information

Introduction: the Semantic Web

- *Social Semantic Web*: RDF data containing social information
- Semantic Web = complex system

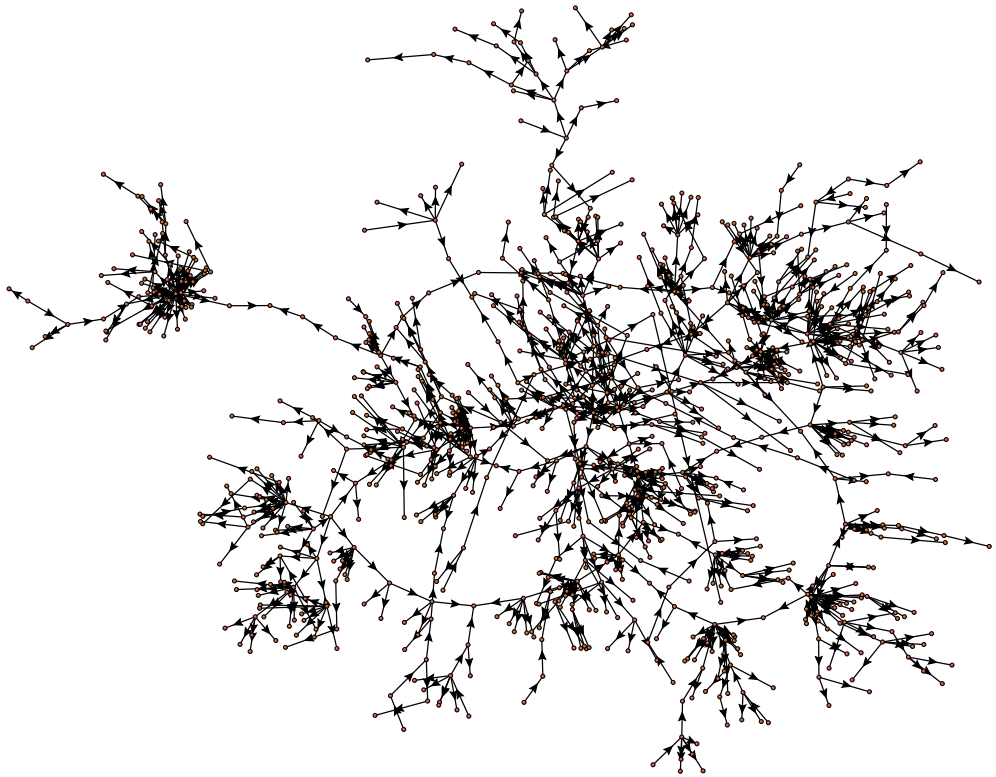
Introduction: the Semantic Web

- *Social Semantic Web*: RDF data containing social information
- Semantic Web = complex system
  - short paths



Introduction: the Semantic Web

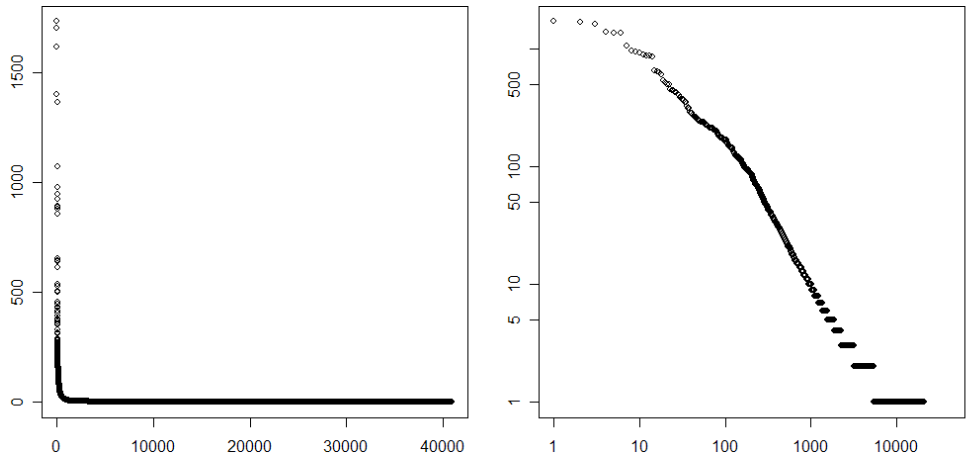
- *Social Semantic Web*: RDF data containing social information
- Semantic Web = complex system
  - short paths
  - clustering



Introduction: the Semantic Web

- *Social Semantic Web*: RDF data containing social information
- Semantic Web = complex system

  - short paths
  - clustering
  - skewed degree distribution: $P(k) \approx Ak^{-\gamma}$



Introduction: Questions

- Social Network Analysis with Semantic Web data
  - prior work by Peter Mika, Li Ding etc.
  - goal: study network properties of multiple aspects $\longrightarrow$ one RDF graph as 'master'

Introduction: Questions

- Social Network Analysis with Semantic Web data
  - prior work by Peter Mika, Li Ding etc.
  - goal: study network properties of multiple aspects $\longrightarrow$ one RDF graph as 'master'
- It is well-known that properties like degree distribution are skewed, but *how* skewed exactly?

Step 1: extraction query

- **Goal**: one ‘master’ graph, multiple derived graphs

Step 1: extraction query

- **Goal**: one ‘master’ graph, multiple derived graphs
- **How?** with SPARQL, query language for RDF

Step 1: extraction query

- **Goal**: one ‘master’ graph, multiple derived graphs
- **How?** with SPARQL, query language for RDF
  - `SELECT` query $\longrightarrow$ result table

Step 1: extraction query

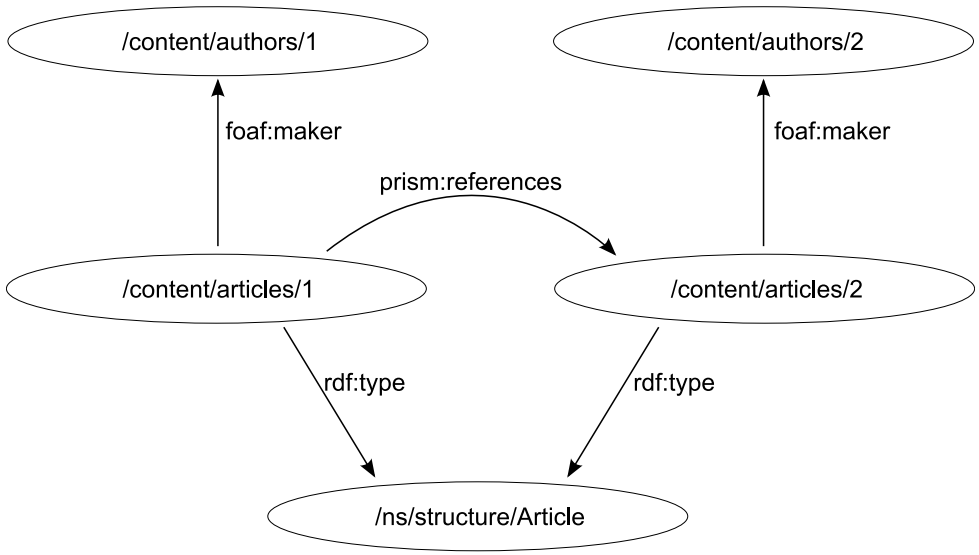
- **Goal**: one ‘master’ graph, multiple derived graphs
- **How?** with SPARQL, query language for RDF
  - `SELECT` query $\longrightarrow$ result table
  - `CONSTRUCT` query $\longrightarrow$ result RDF graph

Step 1: extraction query

- **Goal**: one ‘master’ graph, multiple derived graphs
- **How?** with SPARQL, query language for RDF
  - `SELECT` query $\longrightarrow$ result table
  - `CONSTRUCT` query $\longrightarrow$ result RDF graph
  - `ASK`
  - `DESCRIBE`

Step 1: extraction query

- **Goal**: one 'master' graph, multiple derived graphs
- **How?** with SPARQL, query language for RDF
  - `SELECT` query $\longrightarrow$ result table
  - `CONSTRUCT` query $\longrightarrow$ result RDF graph
  - `ASK`
  - `DESCRIBE`
- **Example**:



```
BASE <http://metastore.ingentaconnect.com>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX prism:
  <http://prismstandard.org/namespaces/1.2/basic>
PREFIX ex: <http://example.com/ns/>

CONSTRUCT { ?author1 ex:cites ?author2 }
WHERE {
  ?art1  a                  </ns/structure/Article> ;
         foaf:maker         ?author1 ;
         prism:references   ?art2 .
  ?art2  a                </ns/structure/Article> ;
         foaf:maker       ?author2 .
}
```

Step 2: secondary graph to SNA format

- This is optional, but. . .

Step 2: secondary graph to SNA format

- This is optional, but... most RDF and FOAF tools are quite limited

Step 2: secondary graph to SNA format

- This is optional, but... most RDF and FOAF tools are quite limited
- Convert RDF to format for (social) network analysis

Step 2: secondary graph to SNA format

- This is optional, but... most RDF and FOAF tools are quite limited
- Convert RDF to format for (social) network analysis
  - Disadvantage: requires new conversion every time something changes in source RDF

Step 2: secondary graph to SNA format

- This is optional, but... most RDF and FOAF tools are quite limited
- Convert RDF to format for (social) network analysis
  - Disadvantage: requires new conversion every time something changes in source RDF
  - through **pyNetConv** to Pajek, GML, GraphML, ...

Unevenness

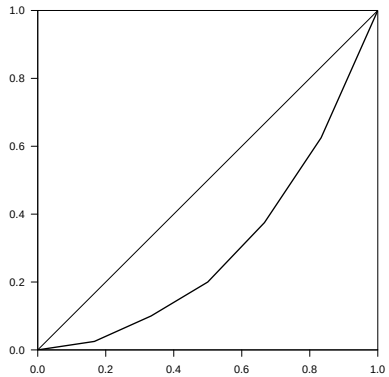
- Or: skewedness, inequality
- Intuitive notion, but how can it be expressed?

Unevenness

- Or: skewedness, inequality
- Intuitive notion, but how can it be expressed?
  - studied in econometrics and informetrics
  - several measures available (not all good)

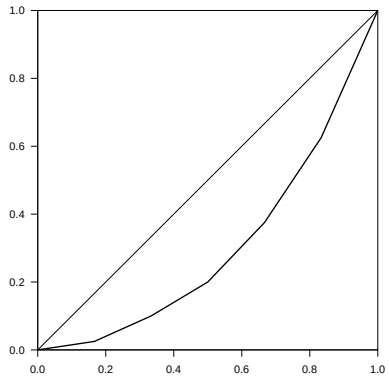
Unevenness

- Or: skewedness, inequality
- Intuitive notion, but how can it be expressed?
  - studied in econometrics and informetrics
  - several measures available (not all good)
- Lorenz curve: *graphical* measure



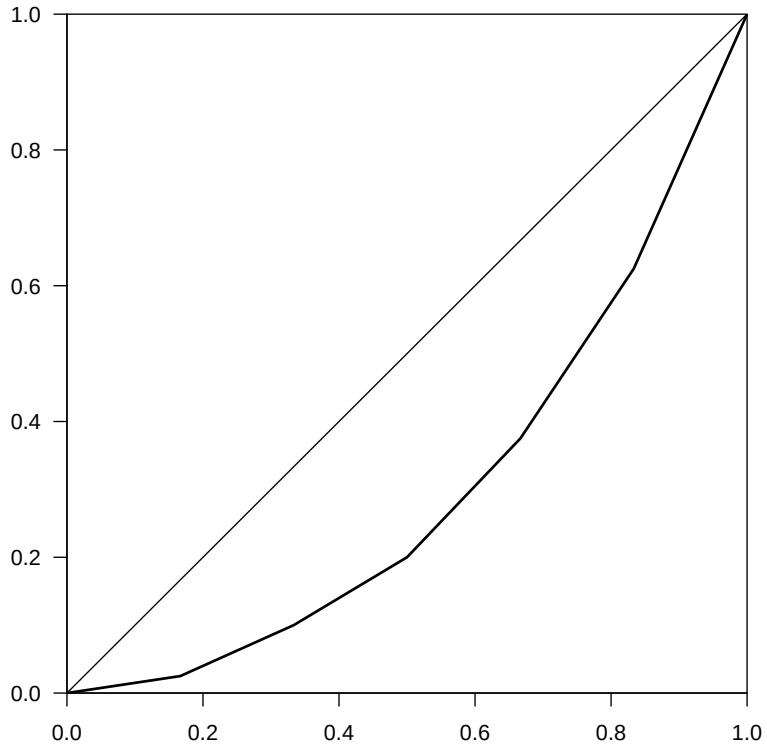
Unevenness

- Or: skewedness, inequality
- Intuitive notion, but how can it be expressed?
  - studied in econometrics and informetrics
  - several measures available (not all good)
- Lorenz curve: *graphical* measure

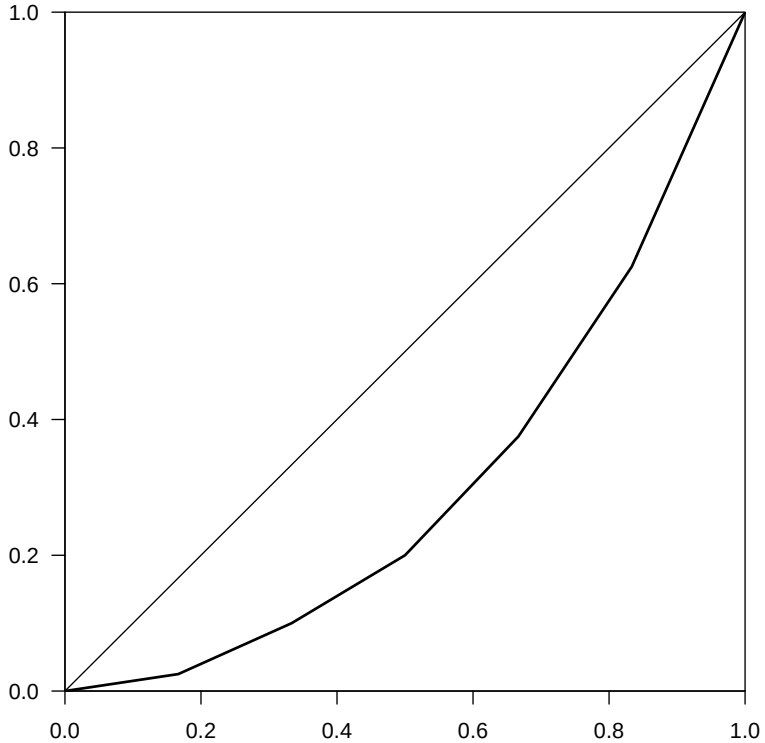


- Gini index: *numerical* measure
  - equivalent to Pratt's measure

- Lorenz curve

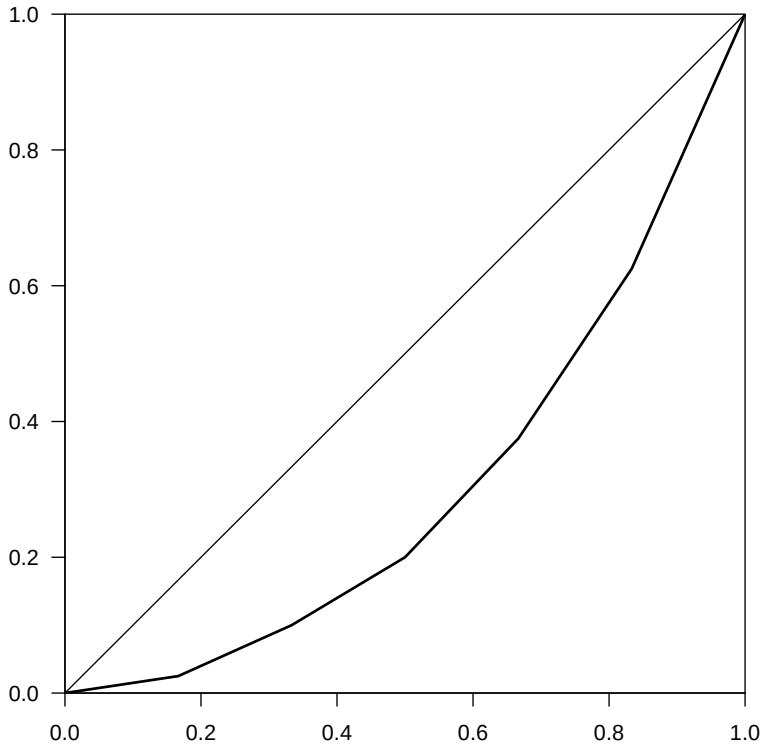


- Lorenz curve



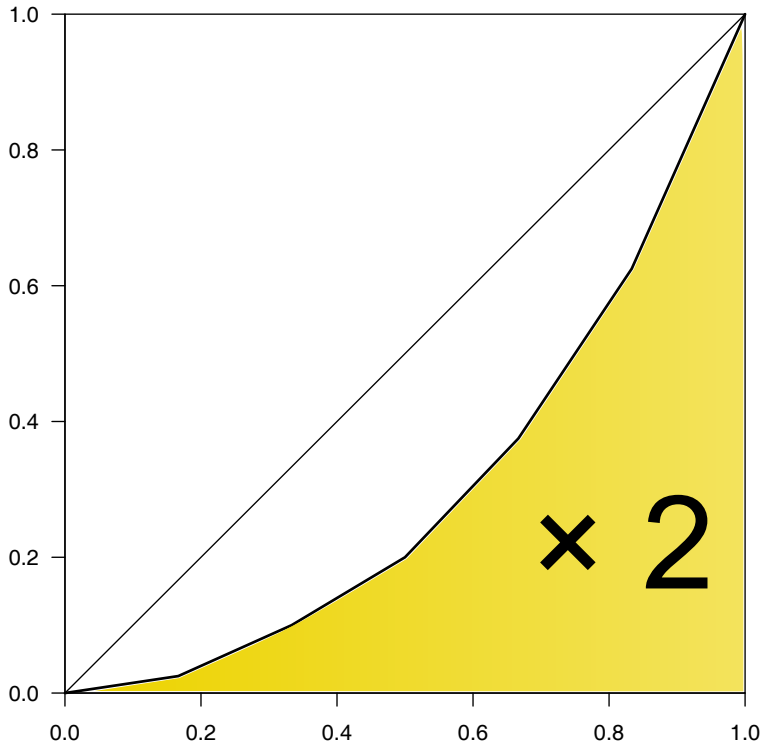
- Gini index: $G'(X) = \frac{2}{\mu N^2}\left(\sum_{j=1}^{N}(N+1-j)x_j\right) - \frac{1}{N}$

- Lorenz curve



- Gini index: $G'(X) = \frac{2}{\mu N^2}\left(\sum_{j=1}^{N}(N+1-j)x_j\right) - \frac{1}{N}$
  - $=$ twice the area under the convex Lorenz curve

- Lorenz curve



- Gini index: $G'(X) = \frac{2}{\mu N^2}\left(\sum_{j=1}^{N}(N+1-j)x_j\right) - \frac{1}{N}$
  - $=$ twice the area under the convex Lorenz curve

Example: Agrippa

- Agrippa = catalogue and database of AMVC Letterenhuis, Antwerp (see `http://museum.antwerpen.be/amvc_letterenhuis/index_eng.html`)

Example: Agrippa

- Agrippa = catalogue and database of AMVC Letterenhuis, Antwerp (see `http://museum.antwerpen.be/amvc_letterenhuis/index_eng.html`)
- Contains information about
  - archival materials (manuscripts, paintings, letters, posters etc.)
  - creators of these materials (people and organizations)
  - relations between these

Example: Agrippa

- Agrippa = catalogue and database of AMVC Letterenhuis, Antwerp (see `http://museum.antwerpen.be/amvc_letterenhuis/index_eng.html`)
- Contains information about
  - archival materials (manuscripts, paintings, letters, posters etc.)
  - creators of these materials (people and organizations)
  - relations between these
- Storage: Jena

Example: Agrippa

- Agrippa = catalogue and database of AMVC Letterenhuis, Antwerp (see `http://museum.antwerpen.be/amvc_letterenhuis/index_eng.html`)
- Contains information about
  - archival materials (manuscripts, paintings, letters, posters etc.)
  - creators of these materials (people and organizations)
  - relations between these
- Storage: Jena
- Queried through: SPARQL protocol

Example: Agrippa

- Agrippa = catalogue and database of AMVC Letterenhuis, Antwerp (see `http://museum.antwerpen.be/amvc_letterenhuis/index_eng.html`)
- Contains information about
  - archival materials (manuscripts, paintings, letters, posters etc.)
  - creators of these materials (people and organizations)
  - relations between these
- Storage: Jena
- Queried through: SPARQL protocol
- Many interesting graphs can be derived

Correspondence in Agrippa

- Example: writers and recipients of correspondence in Agrippa

Correspondence in Agrippa

- Example: writers and recipients of correspondence in Agrippa
- In other words:
  - Nodes = persons (sometimes acting on behalf of an organization)
  - Arcs = letters from writer to recipient

Correspondence in Agrippa

- Example: writers and recipients of correspondence in Agrippa
- In other words:
  - Nodes = persons (sometimes acting on behalf of an organization)
  - Arcs = letters from writer to recipient
- Simple SPARQL query:

```
PREFIX agrippa: <http://anet.ua.ac.be/agrippa#>
CONSTRUCT {
  ?sender <urn:agrext#writesLetterTo> ?recipient
} WHERE {
  ?context agrippa:hasLetterWriter ?sender .
  ?context agrippa:hasRecipient ?recipient .
}
```

Centrality

- **Degree centrality**: number of connections to a node

Centrality

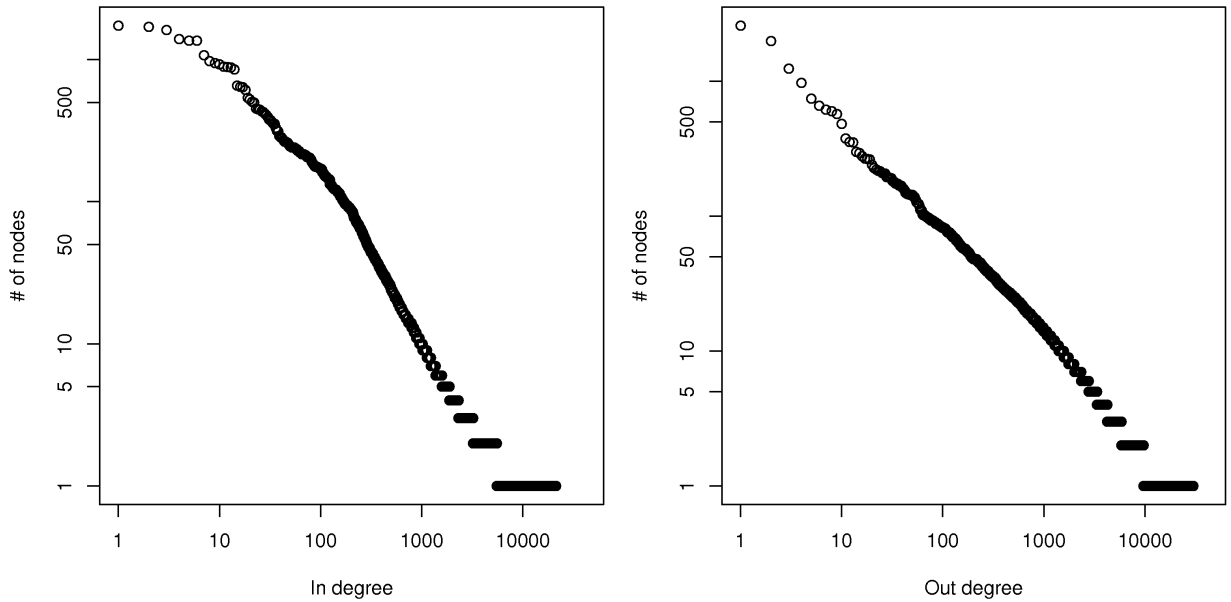
- **Degree centrality**: number of connections to a node
- **Betweenness centrality**: how important is this node for establishing short paths between other nodes?

Centrality

- **Degree centrality**: number of connections to a node
- **Betweenness centrality**: how important is this node for establishing short paths between other nodes?
- **Closeness centrality**: how fast can other nodes be reached from this one?

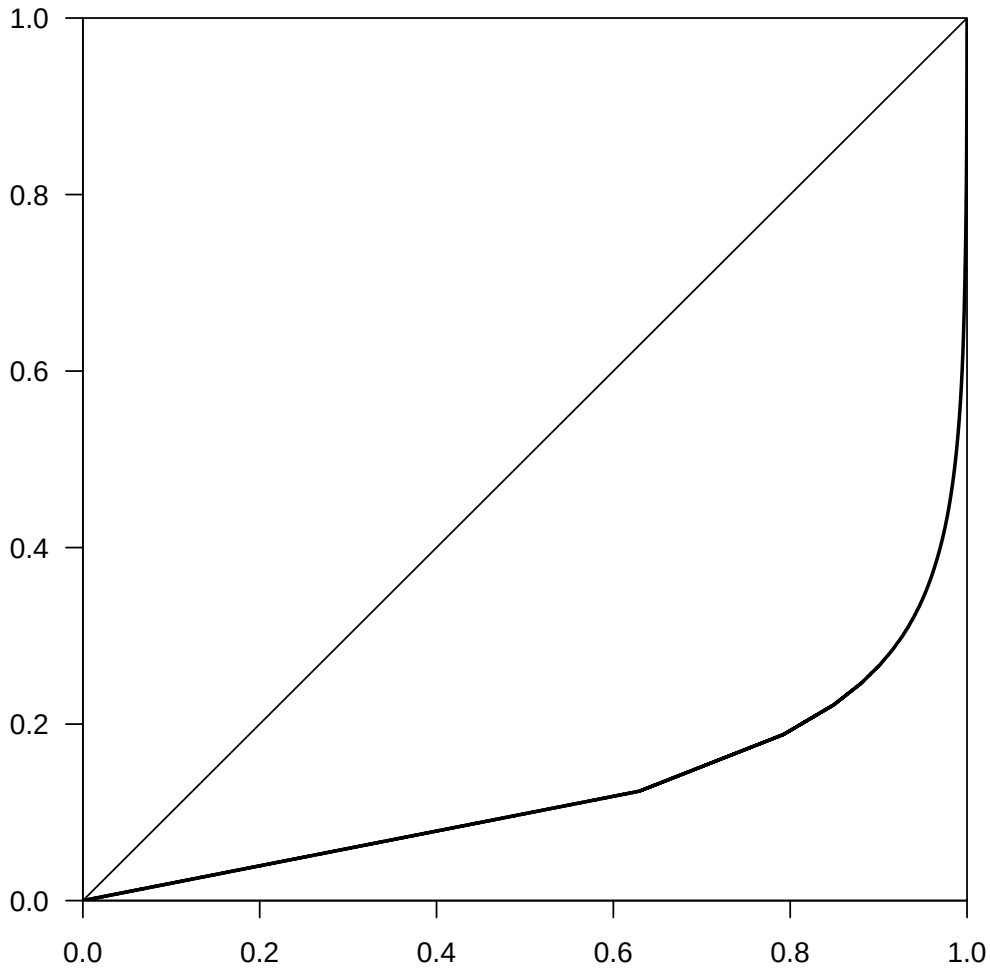
Centrality

- **Degree centrality**: number of connections to a node
- **Betweenness centrality**: how important is this node for establishing short paths between other nodes?
- **Closeness centrality**: how fast can other nodes be reached from this one?
- Example of unevenness (in degree and out degree):



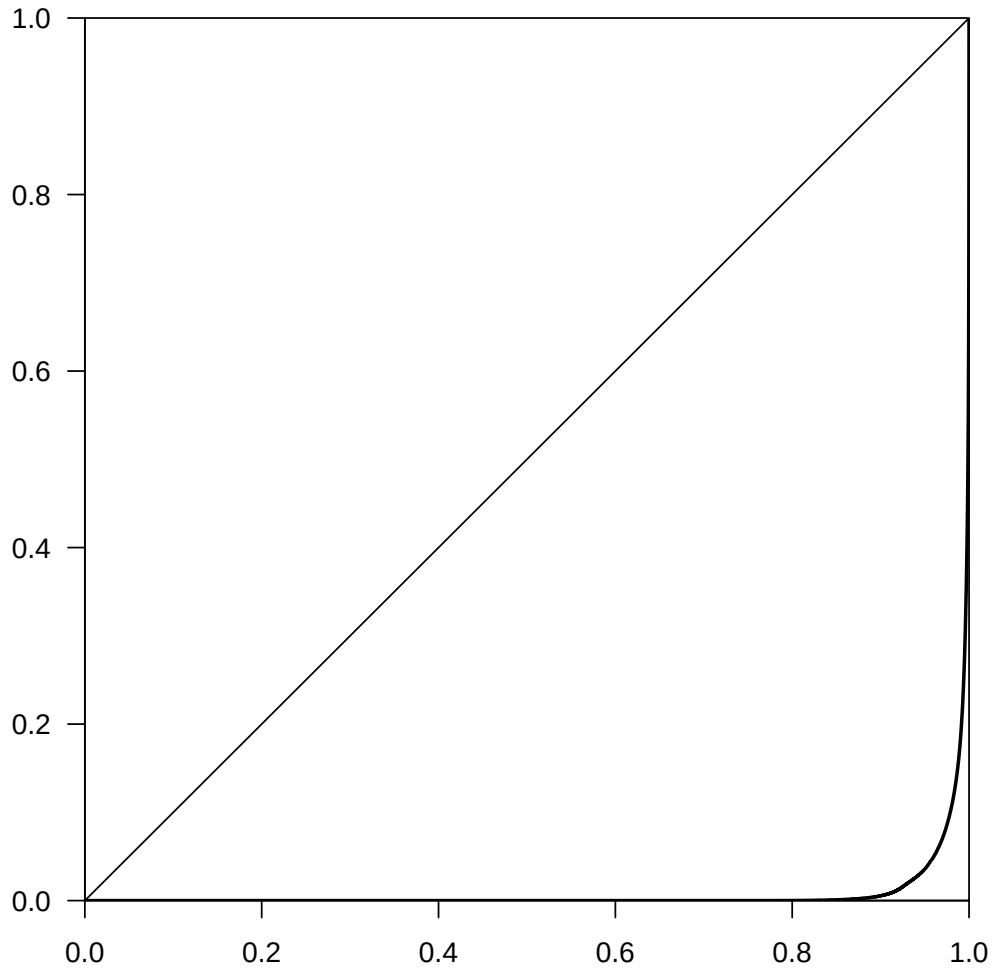
Lorenz curves for centrality measures

Degree centrality (DC)



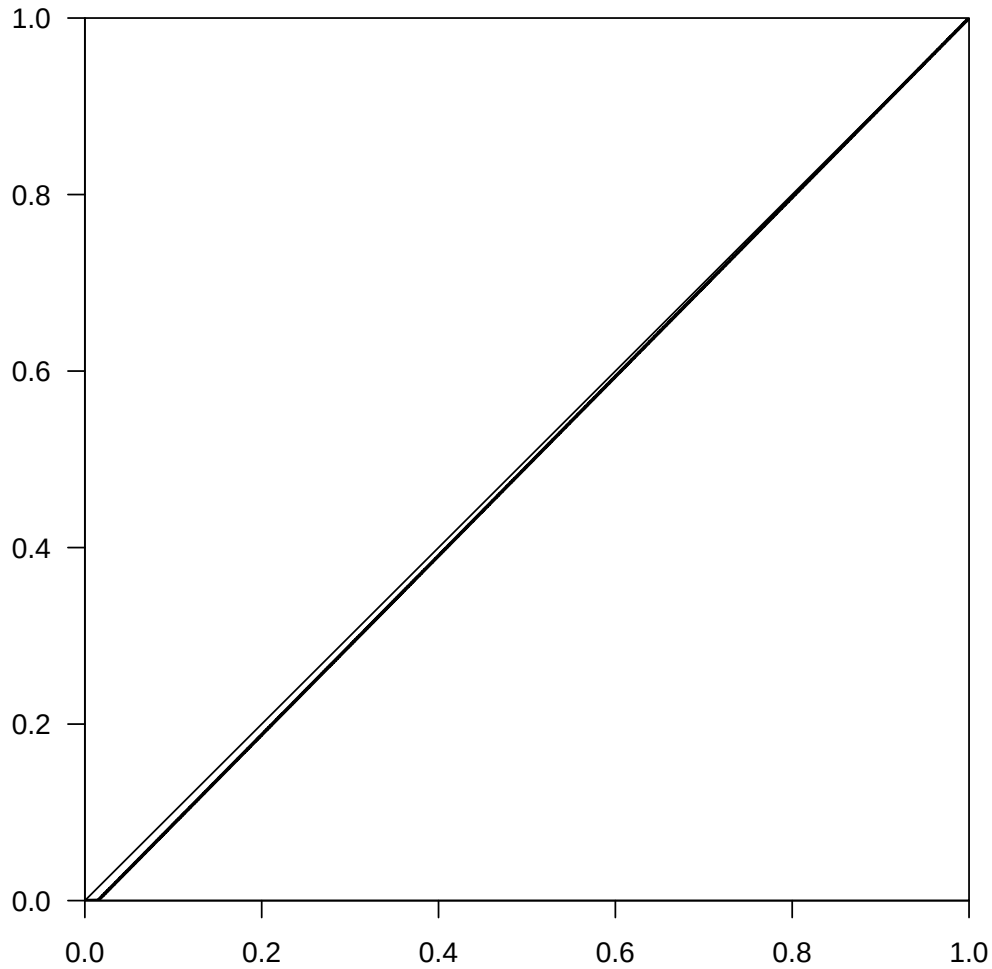
Lorenz curves for centrality measures

Betweenness centrality (BTC)



Lorenz curves for centrality measures

Closeness centrality (CC)



Discussion

- According to Lorenz curve:
  - BTC is more uneven than DC and CC
  - slight overlap between DC and CC

Discussion

- According to Lorenz curve:
  - BTC is more uneven than DC and CC
  - slight overlap between DC and CC
- According to Gini: $G'(BTC) < G'(DC) < G'(CC)$

Discussion

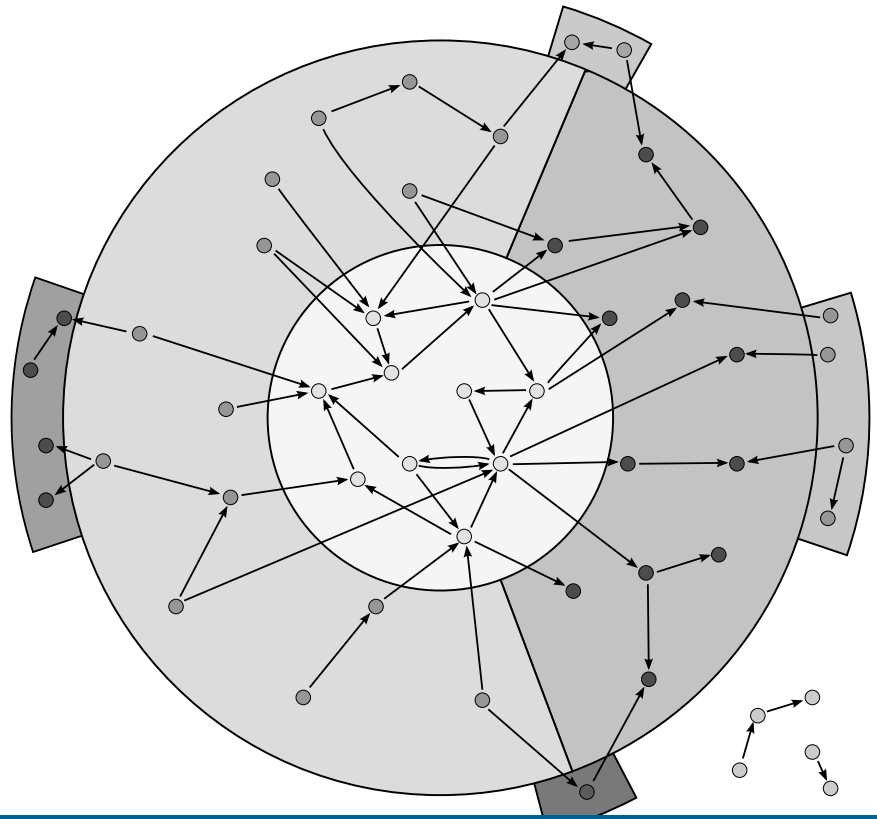
- According to Lorenz curve:
  - BTC is more uneven than DC and CC
  - slight overlap between DC and CC
- According to Gini: $G'(BTC) < G'(DC) < G'(CC)$
- Why such huge differences?

Why such differences?

- CC is quite even, due to the small world effect
  - diameter $D = 11$
  - average length of shortest paths $= 3.85$

Why such differences?

- CC is quite even, due to the small world effect
  - diameter $D = 11$
  - average length of shortest paths $= 3.85$
- BTC is very *un*even due to the bow-tie/corona structure



In summary...

- Simple **two-step methodology** with central place for SPARQL: balance between power and usability

In summary...

- Simple **two-step methodology** with central place for SPARQL: balance between power and usability
- **Unevenness in network measures** can be used to test hypotheses regarding network structure

In summary...

- Simple **two-step methodology** with central place for SPARQL: balance between power and usability
- **Unevenness in network measures** can be used to test hypotheses regarding network structure
- **Future**:
  - testing on other (kinds of) networks
  - predictive power of unevenness?

Thank you!

Any questions?