WWW, Wikipedia and OSNs

DSTA

0.1 Networks of Humans

Theme: no one controls the evolution of the network, which is self-organizing

What is represented (self, news, opinion, concept) and its **lifecycle** determines the structure and the research questions

. . .

look at how they connect and when

Direction of communication is important

```
import networkx as nx

eu_DG = nx.DiGraph()
```

1 Getting data

1.1 WWW

- a Networkx digraph will represent connectivity
- a companion dictionary maps vertices to URLs of the relative pages
- source: a *scrape* of the 2005 ".eu" domain

1.2 Twitter

- supported by the Twython module
- requires Twitter registration/API token
- alternative platforms exist, e.g. Tweety (+NLTK)
- interesting: the network of mentions as a voting system

1.3 Wikipedia

- a network of concepts (lemmas/lemmata) maintained by humans (and some bot)
- time-stamped evolution of the network is available [here]
- contrary to *curated* taxonomies, e.g., [Linnaeus, 1735], this is not a tree

. . .

a **directed acyclic graph** is the reference model

2 Ranking Algorithms: PageRank

2.1 PageRank idea

Assign a **rank** to each vertex (page) on the basis of its *importance* in the navigation of the network.

. . .

Importance will then be captured by the relative value of the dominant Eigenvector of a new matrix P that represents *navigation*

2.2 Variables used

A: directed adjacency matrix (admits *dangling* ends)

. . .

$K0^{-1}$: 0 everywhere but $\frac{1}{k_j}$ on the main diagonal

. . .

$N = A \cdot K0^{-1}$

. . .

E: 0 everywhere but $\frac{1}{|V|}$ on the main diagonal

$$P = \alpha N + (1 - \alpha)E$$

Experimentally, set $\alpha = 0.85$

I.e., $1 - \alpha$ times navigation will *jump out* of a path and into an arbitrary *restart* node.

3 Ranking Algorithms: HITS

3.1 HITS idea

Hyperlink-Induced Topic Search [Kleinberg, 1999]

Sees importance of a node in a more nuanced way:

Pages that are important for consultation, e.g., train schedules, have *authority* and tend to be *terminal*

. . .

Well-connected *hub* pages that facilitate navigation, e.g., Time Out, are useful but not authoritative per se

. . .

1. authority score $\mathbf{au}(\mathbf{i})$
2. hub score $\mathbf{h}(\mathbf{i})$

3.2 HITS as Mutual recursion

Hub-iness influences authority which in turns influences hub-iness:

$$au(i) \propto \sum_{j \to i} h(j)$$

page i is authoritative proportionally to the sum of the hub-iness of the pages that link to it.

$$h(i) \propto \sum_{i \to j} au(j)$$

page i is hub proportionally to the sum of the authoritativeness of pages that it links to.

3.3 Computing HITS scores

We could start with assigning 1 everywhere and hoping that mutual recursion will converge to stable au and h values.

As with Von Mises' method, we normalise vectors to 1 at each iteration.

3.4 Linear Algebra derivations

$$\mathbf{h} \propto AA^T\mathbf{h} = \lambda_h AA^T\mathbf{h}$$

$$\mathbf{au} \propto A^TA\mathbf{au} = \lambda_{au} A^TA\mathbf{au}$$

I.e., we can find **h** and **au** separately by solving the eigenvalue problem for the matrices AA^T and A^TA

3.5 Main result

For *primitive* matrices (i.e., connected networks, no dead-ends/sinks)

$$\mathbf{h} \propto AA^T\mathbf{h} = \lambda_h AA^T\mathbf{h}$$

$$\mathbf{au} \propto A^TA\mathbf{au} = \lambda_{au} A^TA\mathbf{au}$$

- convergence is assured;
- dominant λ is unique and
- values for **h** and **au** will be all positive, as desired.

(negative values have no interpretation here)

4 Community detection

4.1 Finding social structures

this is an example of Provost-Fawcett's problems

- 4: Clustering
- 5: co-occurrence grouping

. . .

For homogeneous networks, eg., country-to-county of Ch. 2

Community: nodes that are closely connected with each other by *strenght* or *density*

Resolution limit: communities with less than $\sqrt{|V|}$ members cannot be properly identified.

4.2 Givan-Newman

1. Rank edges by their *help to connectivity*
2. remove the top-ranking edge
3. repeat until loss of connection
4. now-isolated areas are called communities

Hyp: Betweenness centrality captures *help to connectivity*

5 Modularity

5.1 As an optimization prob.

Istance: an adj. matrix A, a small integer g

Solution: a partition of V into g *groups*

. . .

Measure: maximise Q: the overall modularity measure

Interpretation: how likely is a random walker to leave the community?

5.2 The Q factor

Let $E_{g\times g}$ be the cross-group matrix and f_i the sum of col. i

. . .

Electrical conductance:

$$Q = \sum_{i=1}^{g} e_{ii} - f_i^2$$

. . .

Complexity: NP-complete

Even random networks might exibit densifications that might look as c.