Implicit translation of Wikipediæ via Random Graph Embeddings

David J. Marchette

Naval Surface Warfare Center
david.marchette@navy.mil

August 18, 2009

Graphs vs Networks

- A graph is a pair (V, E) of vertices and edges.
  - V is a set (called vertices or actors).
  - E is a set of pairs of vertices (unordered, if a graph, ordered if a directed graph).
  - We'll assume the graph is simple: no self loops or multiple edges.
- A network is a graph with extra information:
  - Weights or covariates on the:
    - Edges.
    - Vertices.
    - There may be external information about the graph as a whole.
  - The usual definition of a network is a directed graph with weighted edges, but we will broaden this definition.

Framework: Setup

- Here is a basic framework that is general enough to discuss the various issues.
- Given $D = (G, H)$ where
  - $G = ((V, W_G), E_G)$ is an unweighted simple graph, $|V| = n$, $|W_G| = m_G$.
  - $H = ((V, W_H), E_H)$ is a unweighted simple graph, $|W_H| = m_H$.
  - $N = |V| + |W_G| + |W_H| = n + m_G + m_H$.
- Note that we assume the subset V of the vertex sets are the same (and static) in both graphs.
- The vertex sets W_G, W_H are vertices for which the correct association (if any) is unknown.
- Various extensions are possible, but this is enough to illustrate some ideas.

Framework: Embedding

- We seek to embed D in some space in which we can perform inference.
- The inference we are primarily concerned with is that on the unknown vertex sets W_G and W_H: we want to find associations among these and/or the known vertices V.
- Thus, the embedding $h : \mathcal{D} \rightarrow \mathbb{R}^{N \times d}$ is an embedding of the N vertices.
- Ideally, we want the best such embedding for a given inference task, measured by β:

$$h^* = \arg\max_h \beta(h; \mathcal{D})$$

- β will be a measure of how "good" the matching is for the unknown vertices.

Framework: Embedding

- We can embed each piece of D separately (into the "same" $\mathbb{R}^d$:

$$\begin{aligned} G &\mapsto X_G \in \mathbb{R}^{(n+m_G)\times d} \\ H &\mapsto X_H \in \mathbb{R}^{(n+m_H)\times d} \end{aligned}$$

- Then combine the embeddings:

$$(X_G, X_H)' \subset \mathbb{R}^d.$$

- A problem with this approach is to ensure that the embeddings are commensurable. E.g., we may want to perform Procrustes on the embeddings of the V vertices to match them up.

Framework: Embedding

- Or we can embed the totality:

$$(G, H) \mapsto \mathbb{R}^{N \times d}.$$

- In principle, this approach should be superior to separate embedding, in situations in which there is correlation, or similar information, in the different graphs, which should be the case if the unknown vertices do have associations.
- We seek methods which can be applied to the pieces separately or the combination.

Framework: Covariates

- In the application we are interested in, implicit translation of documents, there are covariates of interest (the words in the documents).
- We can (should) incorporate this information, particularly if we have available dictionaries to translate terms.
- In this work we will assume that dictionaries are unavailable, and will not use the language information.
- However, clearly this information should be used, and we will discuss some ideas at the end of the talk.

Implicit Translation

- Translation of a document from one language to another involves the translation of words (phrases, sentences).
- Care must be taken to ensure that the meaning is maintained, while also ensuring that the syntax in the target language is correct.
- Traditionally, implicit translation is the extraction of the basic meaning from a document, and its presentation in a useful form (such as a modern exposition of a Greek mathematics text).

Implicit Translation

- We use *implicit translation* to mean the association of a document in one language to one in the target language that is “the same” in some measurable way:
  - A rough (not word-for-word) translation of the document.
  - A document written on the same topic.
  - A document on a similar topic.
- In our experiments we will use a human-provided “true” association for evaluation.

Random Embedding

- Assign to each vertex v a vector $U_v \in \mathbb{R}^d$.
- Associate to each edge $e = vw$ the vector $g(U_v, U_w) \equiv g(e)$ (for example $g(U_v, U_w) = U_v * U_w$ where multiplication is term-by-term). If the edge is weighted, g operates on the weight as well (for example by multiplication).
- Let $s(v)$ be a subgraph containing v (for example, edges incident on v), and $E_{s(v)}$ its edge set.
- Associate to each vertex v the vector

$$X_v = \frac{1}{|E_{s(v)}|} \sum_{e \in E_{s(v)}} g(e).$$

- So, each vertex gets mapped to the centroid of its "associated" edges.

Random Embedding

- U_V can be random or it can correspond to (a function of) the covariates, or other external information.
- The method extends easily to hypergraphs, where a hyper-edge is projected to the product of the vertices' vectors, and the vertices are projected to the average of the projected hyper-edges (Hohman, JSM).
- Note that for some applications, the ability to embed both vertices and edges in the same space may be important.

Wikipediæ

- Wikipedia has versions in over 200 languages.
- Each Wikipedia consists of:
  - A graph of articles (vertices) with links (edges) to other articles.
  - Links to external information (web pages, images) (not used in this work).
  - Text (content of the articles), title.
  - A link to “the same” article in (many) different languages.
- We will use the graphs and the between-language associations in this work.

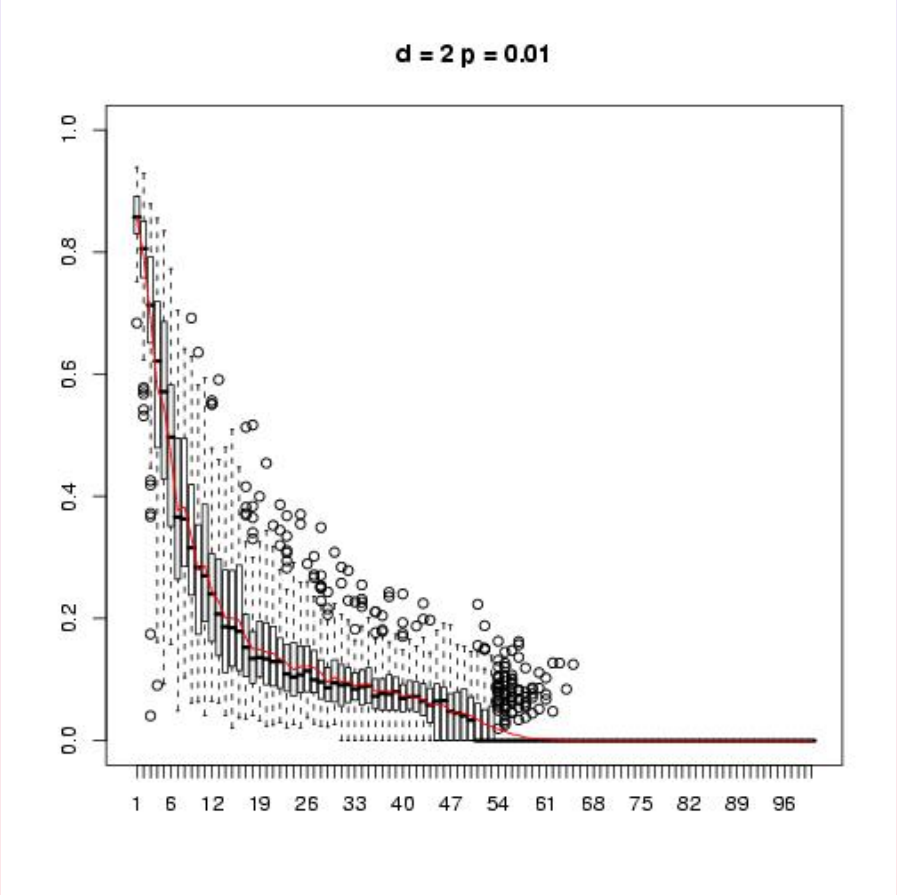
Simulation

- Generate an Erdös-Renyí random graph $G \sim ER(100, p)$, $p \in [0.01, 0.05]$.
- H_k constructed by removing (at random) k edges from G and inserting (at random) k edges not in G.
- So, $G = (V, E_G)$ and $H_k = (V, E_k)$ with $|E_G| = |E_k|$ and the edge sets differ by $2k$ edges.

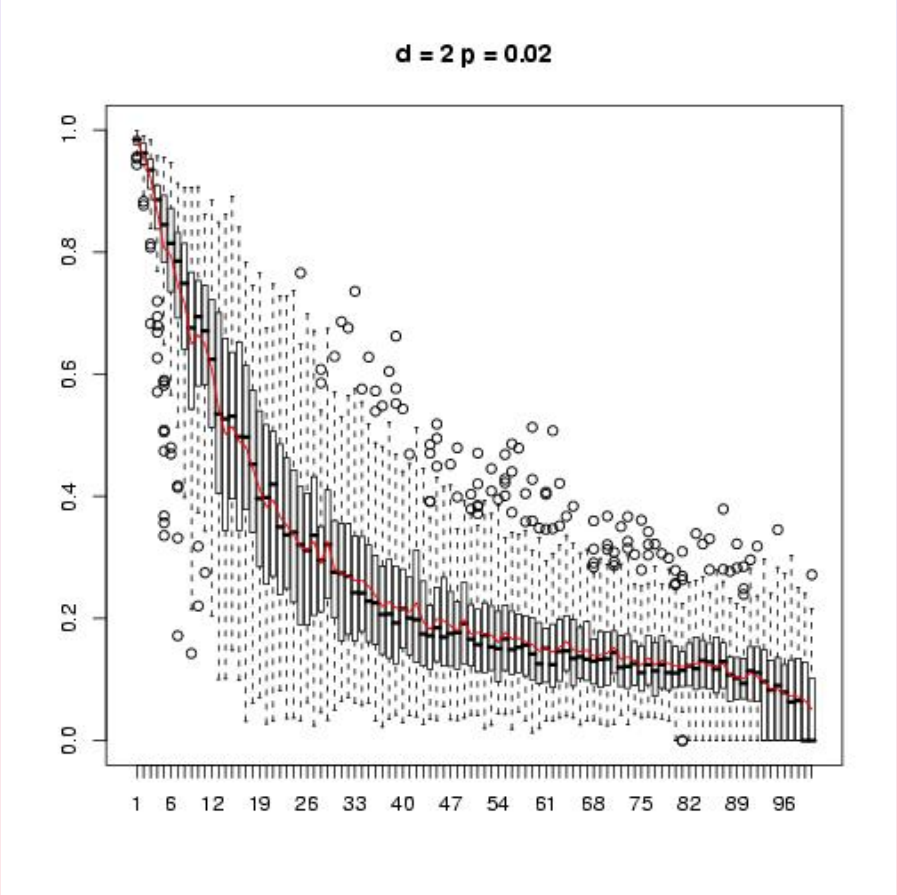
Simulation

- Random embedding performed on G and H:
  - each vertex v is projected to X_v as an element of G and to Y_v as an element of H.
- For each $v \in V$, compute the distance $d(X_v, Y_v)$ and for $\alpha = 0.05$ set the critical value c_α such that $100\alpha\%$ of the distances are less than c_α.
- Report β, the proportion of $d(X_v, Y_w) > c_\alpha$, for $v \neq w$.
- 100 simulations per set of parameters.

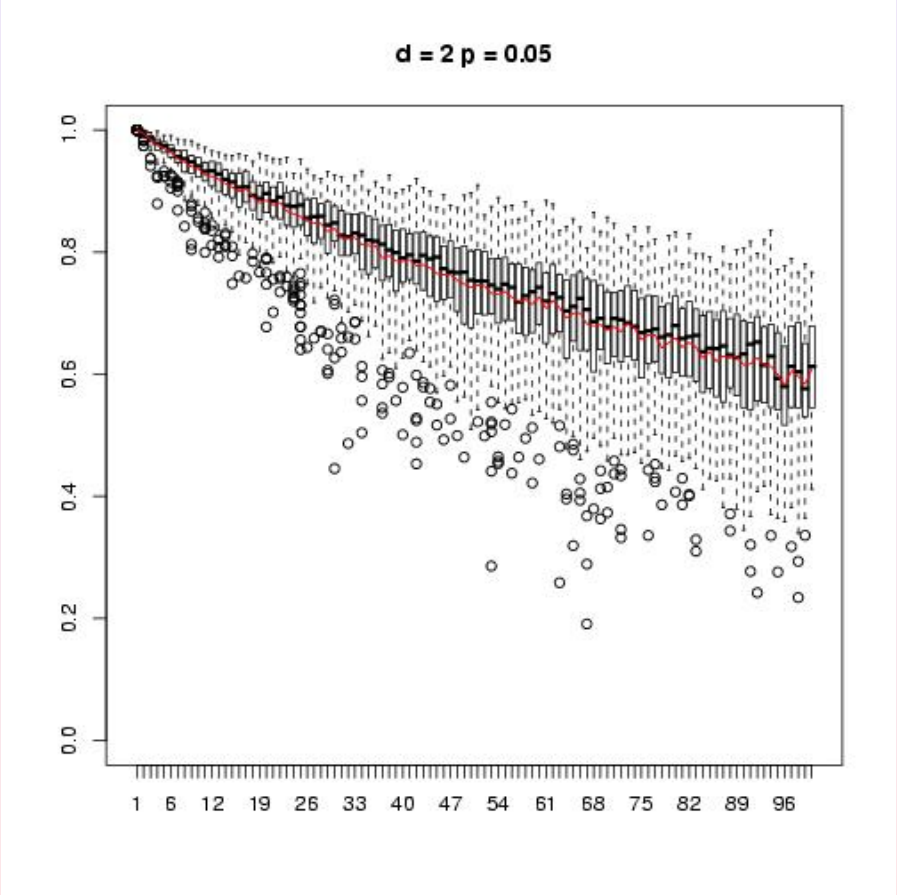
$d = 2$



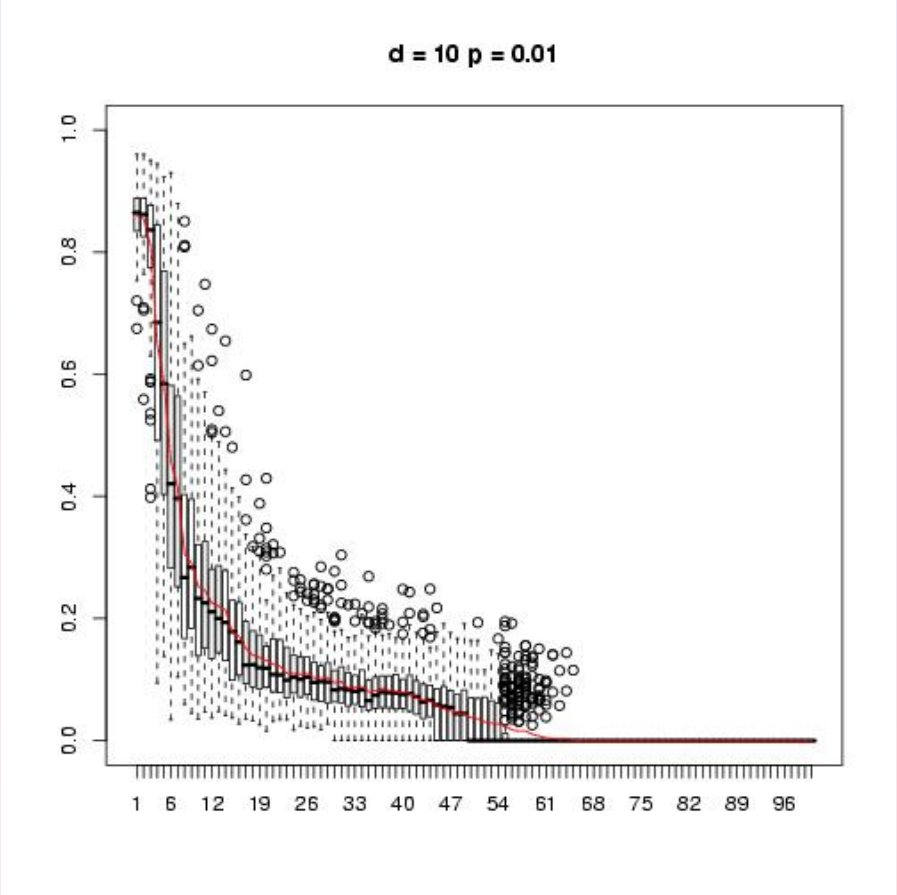
$d = 2$



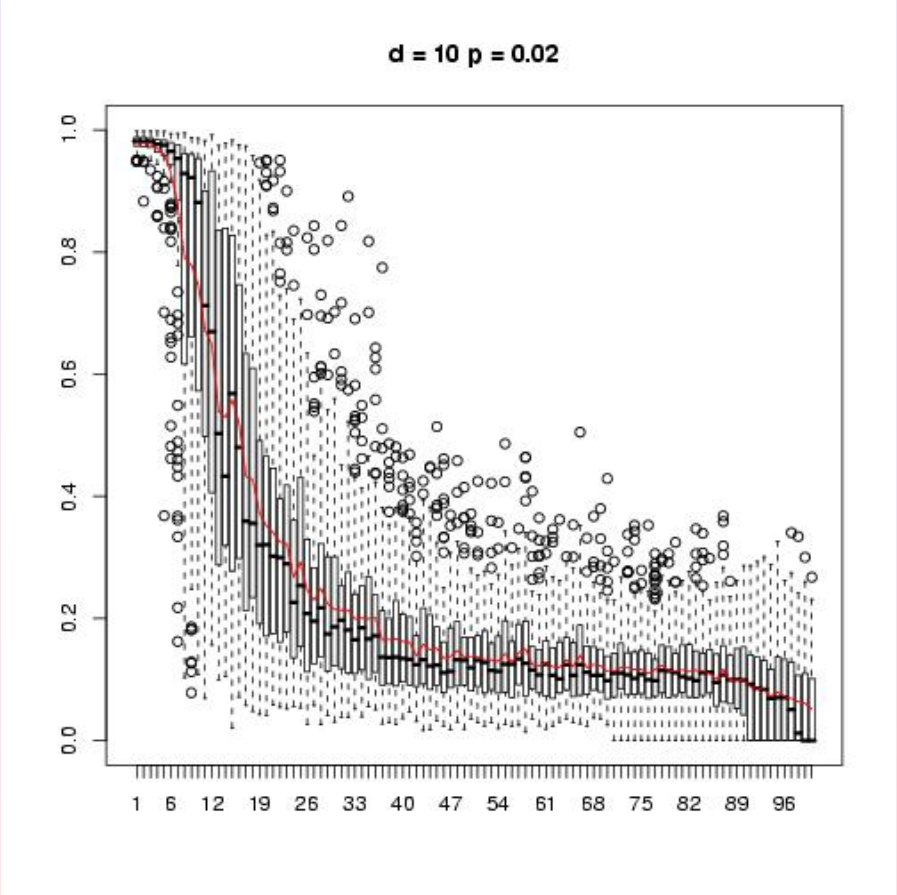
$d = 2$



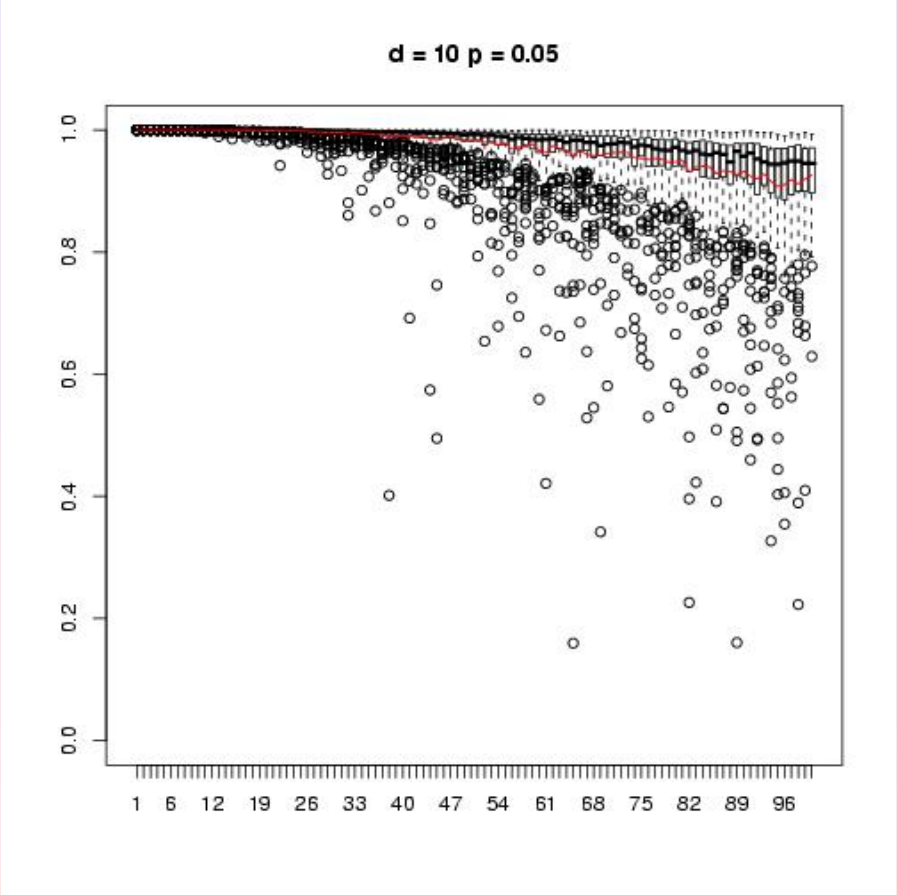
$d = 10$



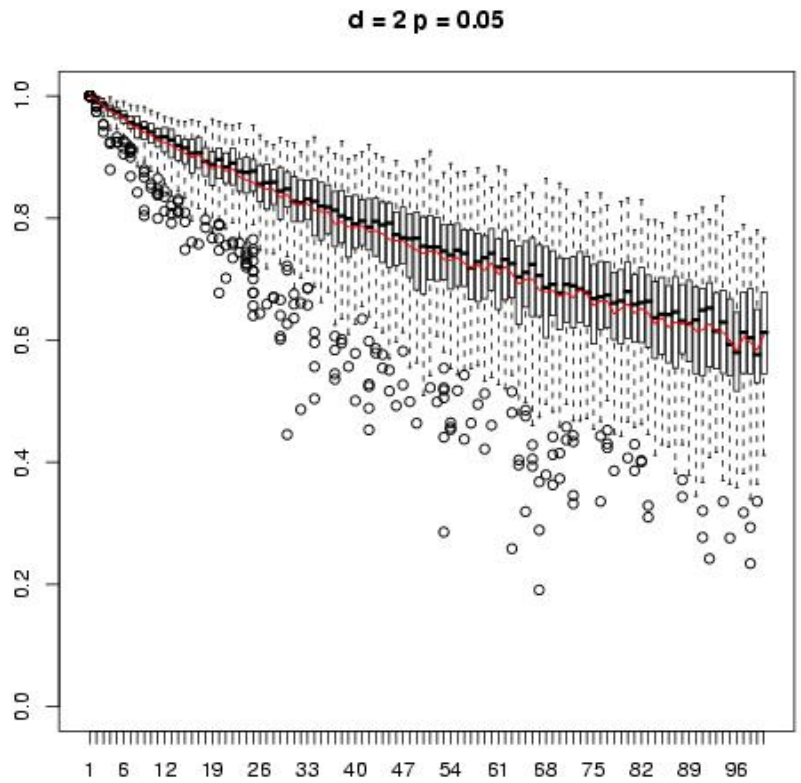
$d = 2$

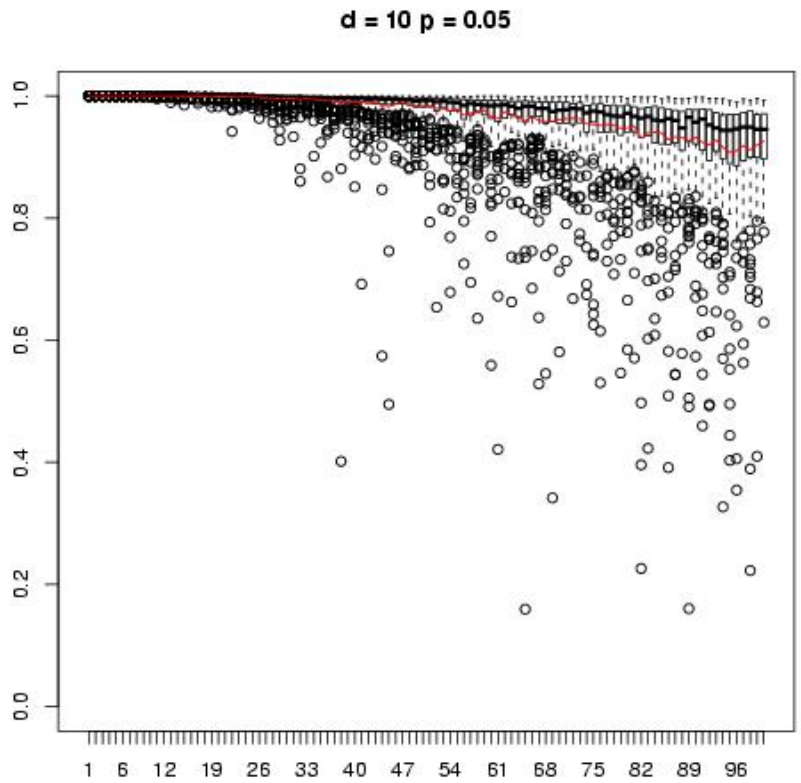


$d = 2$



$d = 2$ vs $d = 10$

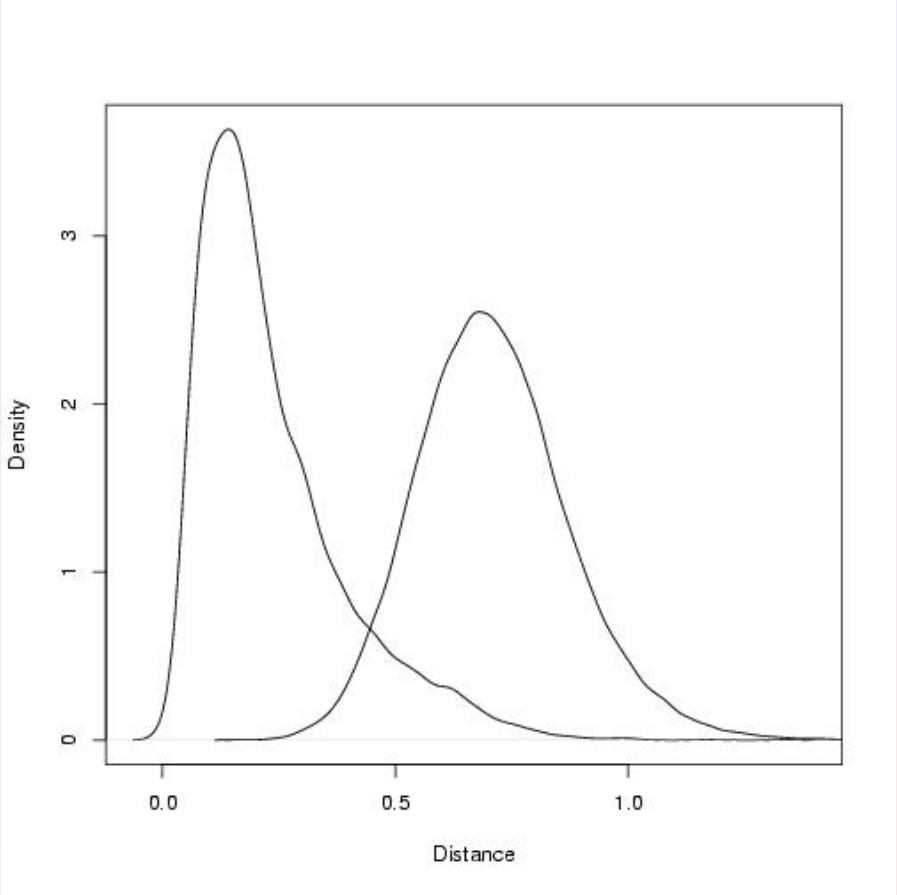

d = 2 p = 0.05
0.0
0.2
0.4
0.6
0.8
1.0
1 6 12 19 26 33 40 47 54 61 68 75 82 89 96


d = 10 p = 0.05
0.0
0.2
0.4
0.6
0.8
1.0
1 6 12 19 26 33 40 47 54 61 68 75 82 89 96

Afrikaans vs Dutch Results

- $9,136$ Afrikaans articles.
- $322,728$ Dutch articles.
- Project graphs, compute distances between paired and non-paired documents.
- Plot the densities.

Afrikaans vs Dutch



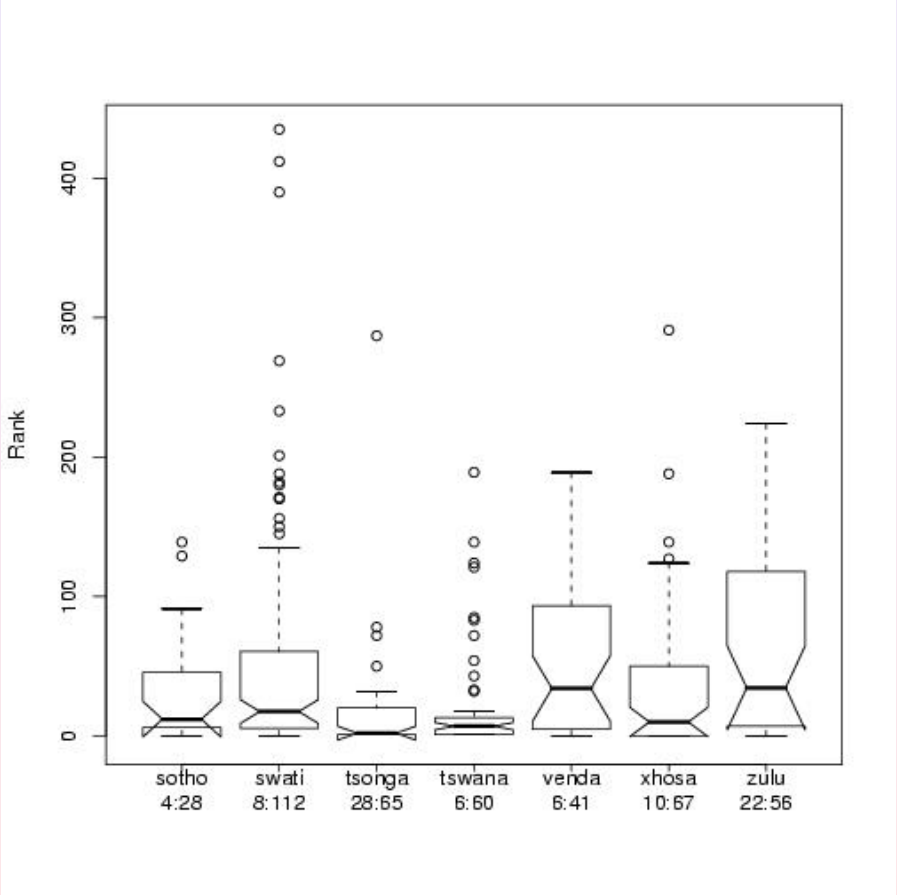
Bantu Statistics

Language	Number of Pages
Southern Sotho	28
Swati	112
Tsonga	65
Tswana	60
Venda	41
Xhosa	67
Zulu	56

Bantu Results

- First, the Afrikaans articles were embedded.
- Each Bantu language was processed separately.
- Each article was projected as the average of its neighbors (leave-one-out crossvalidation, as it were). The projection of the neighbors is the same as used for the Afrikaans.
- The distance between the Bantu article and its paired Afrikaans article was reported (rank among the rest of the distances).

Bantu Results



Discussion

- The simulations show that the random projection maps paired vertices close, provided the graphs are similar.
- Graphs containing very different information (different relationships) must be handled via different means.
- The Wikipedia results show that the method has merit for real-world data.
- Small graphs, such as the Bantu Wikipediæ, are likely to present problems, due to the low overlap of similar topics.
- The random projection is very fast, and can be used for (essentially) arbitrarily large graphs.
- The selection of the proper value for d (embedding dimension) is an issue for future research.
- Comparison with spectral methods is another area of research.

Discussion

- Out-of-sample embedding is an issue:
  - Embed as the average of neighbors (as in Bantu experiment) or similar approach.
  - Choose the embedding to minimize the distance between labeled pairs.
  - Use random embedding as a start for a fast MDS embedding, then use out-of-sample embedding.
- Also need to investigate using the language information.
- See work on fusing disparate information (Priebe, Solka, djm).
- See also Yancey talk on identifying out-of-language documents.