I.  **Introduction: Semantics and graph theory**

A. Scale-free networks of hubs plus nodes with fewer connections:

1. Connection frequencies fit a power law distribution [Barabasi]

2. Describes a number of phenomena represented as a graph

a. Examples

C. Semantic predications representing the assertions in a coherent body of text can be arranged to form a graph with a scale-free node distribution.

B2. We can analyze the characteristics of semantic relationships occurring in such a graph.

a. There are ontological constraints on the paths representing predications (e.g. drugs treat diseases; diseases don’t treat drugs)

C. In this paper, we intend to quantitatively analyze the distribution of the ontological features of a graph of semantic predications as a first step in elucidating the scale-free character of such a graph.

B. C.

II. **Background**

A. Graph theory

1. Define degree, degree centrality,

2.and other graph theory concepts that should be included

B. Source of semantic predications for this study

1. MEDLINE

2. Ontology (basic characteristics of UMLS Metathesaurus and Semantic Network)

3. SemRep

a. How it works

4. General characteristics of semantic predications

a. Semantic predications are ontologically restricted. Arguments are defined by the set of relations in the ontology. Arguments are limited to ontological concept classes that can be joined by specific relations.

C. Graphs representing semantic predications

1. Predications are represented by directed paths of length one

1. This analysis could reinforce the hypothesis that semantic characteristics can explain why networks form as they do.

III. **Methods**

A. We use graphs of semantic predications extracted from text from the biomedical domain. We start with all predications, then limit to different central concepts.

B.The first step is to analyze the quantitative semantic characteristics constituting a hub (a central concept) and all paths of length one emanating from the hub. Such hub-centered paths represent the semantic predications in which the concept of the hub is one of the arguments.

{From here on down needs to be discussed}

A. For four central concepts of different classes…

1. We choose four concepts with different semantic types, and that are distinct within the MeSH hierarchy

2. For each concept, we pull every semantic predication from the SemMed database that includes it as a subject or object (using EDAT), within EDAT values 2002 – 2012 (or all predications from these years, using the EDAT timestamps). We aggregate predications by month.

3. For each instance (or each month), we count (a) frequency of the central concept, (b) number of unique opposing arguments, (c) number of unique predicates, and (d) number of unique semantic types.

4. We use k-means clustering, with 4 centroids (because we have four concepts) to see if instances are unique enough to be clustered by class (i.e., concept). High clustering accuracy indicates that these quantitative values, when combine and aggregated by month, are distinctive to their given concept, when compared with those of other concepts. They are like a fingerprint, of sorts.

B. For central concepts within the same class…

1. For each of the four original concepts, we choose four more concepts that are descendants of the original, according to the MeSH hierarchy, and that share the same semantic type with the original concept.

2. We repeat the steps above, and analyze the clustering accuracy. High accuracy is indicative of distinctive values, as described above. Low accuracy indicates that the four sibling concepts are not as distinctive in terms of the quantitative values we gathered.

C. For the above experiments that yielded high accuracy, we use the data to train a multi-perceptron and test it with new data from another year span

1. To test if a given “fingerprint” is sufficient to identify data with another timestamp

2. Further tests idea that quantitative semantic values contribute to network formation