Hiding Virtues of Ambiguity

Watermarking of Natural Language Text Through Synonym Substitutions

Umut Topkara Mercan Topkara Mikhail J. Atallah

Department of Computer Sciences
Purdue University
West Lafayette, 47906, IN, USA

PURDUE UNIVERSITY

CERIAS

Problem and Key Idea

- **Designing a resilient, practical and easy-to-use watermarking system for natural language text**
  - How can you be sure that your articles/ papers/ blogs/ e-mails are not re-used?
  - Need a computationally light detection process (not AI complete)
  - Adversary can foil string matching
- **Using *robust* synonym substitution for natural language watermarking**
  - We favor more *ambiguous* alternatives (i.e. homographs)
    - smart ⟶ bright
  - The resilience stems from the fact that
    - the adversary does not know *where* the changes were made
    - automated disambiguation is a major difficulty

What is Natural Language Watermarking?

- **Enable copyright holders to enforce their intellectual property ownership on text**
- **Value of Text:**
  - Meaning
  - Grammaticality
  - Style
- **Mark the text such that:**
  - The marking modifications do not reduce text's value
  - Adversary will reduce text's value to remove the mark

Natural Language Challenges

- **Short documents**
- **Low embedding bandwidth**
  - Small number of alternative forms
- **Not all transformations can be applied to a given sentence** **(John went to school.)**
  - Grammar (School was gone by John.)
  - Vocabulary (John matched to school.)
  - Style and fluency (John didn't not go to school.)

PURDUE UNIVERSITY

Natural Language Challenges

- **Powerful Adversary**
  - Can *automatically* edit individual sentences
  - Can permute sentence order
  - Can delete or insert sentences
  - Has access to the same data and software resources

Previous Approaches

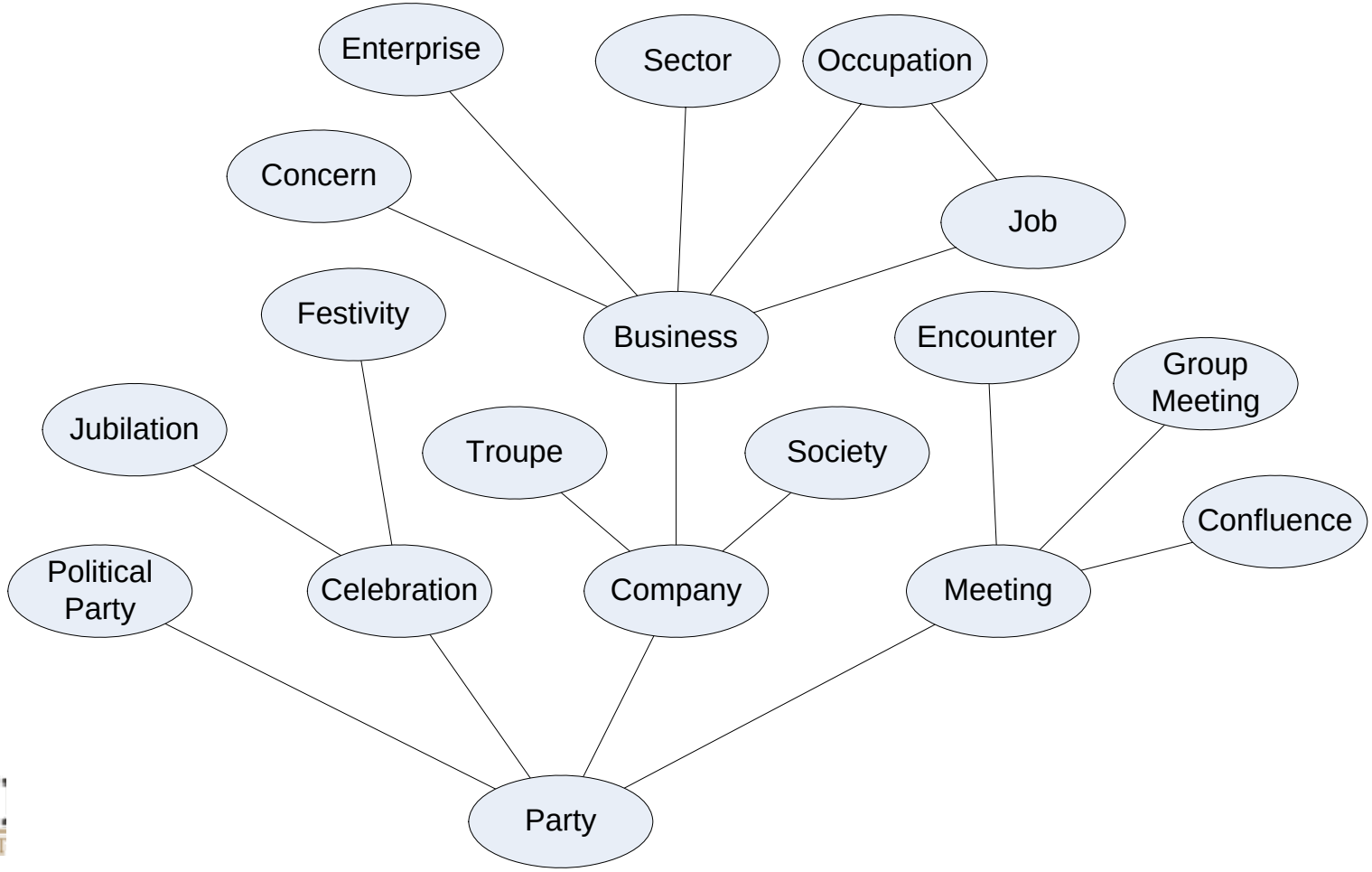
- **Generating the cover text (only for Steganography)**
  - Passive Warden
  - Cover text has no “value”
    - Spammimic (M. Chapman and G. Davida, 2002)
- **Modifying a given cover text**
  - Active Warden
  - Proposed for steganography as well as watermarking

Equimark

- **Performs robust synonym substitution**
  - Ranks alternatives for substitution according to their ambiguity
- **Quantifies the distortion**
  - Keeps the distortion on the original text below a given threshold
  - Restricts the flexibility of the adversary while modifying the watermarked text
- **Does not require either the original text or word sense disambiguation at detection**
- **Follows Kerckhoff's principles**

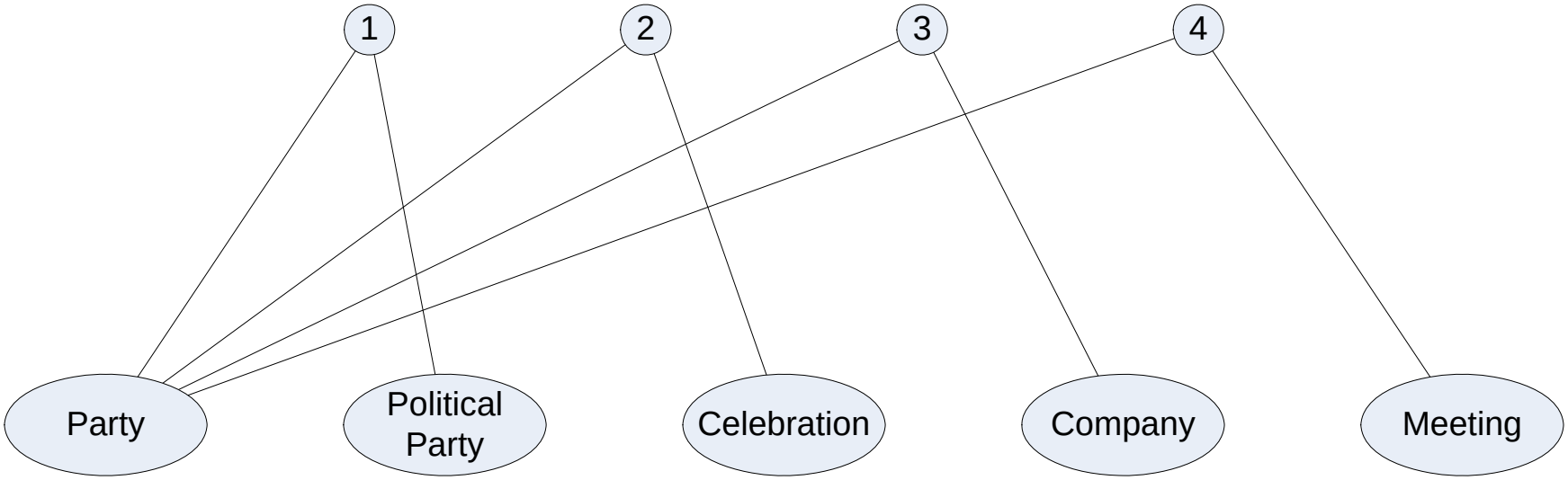
Equimark: Embedding

- **Build a graph, G, of (word, sense) pairs**
  - WordNet



Equimark: Embedding

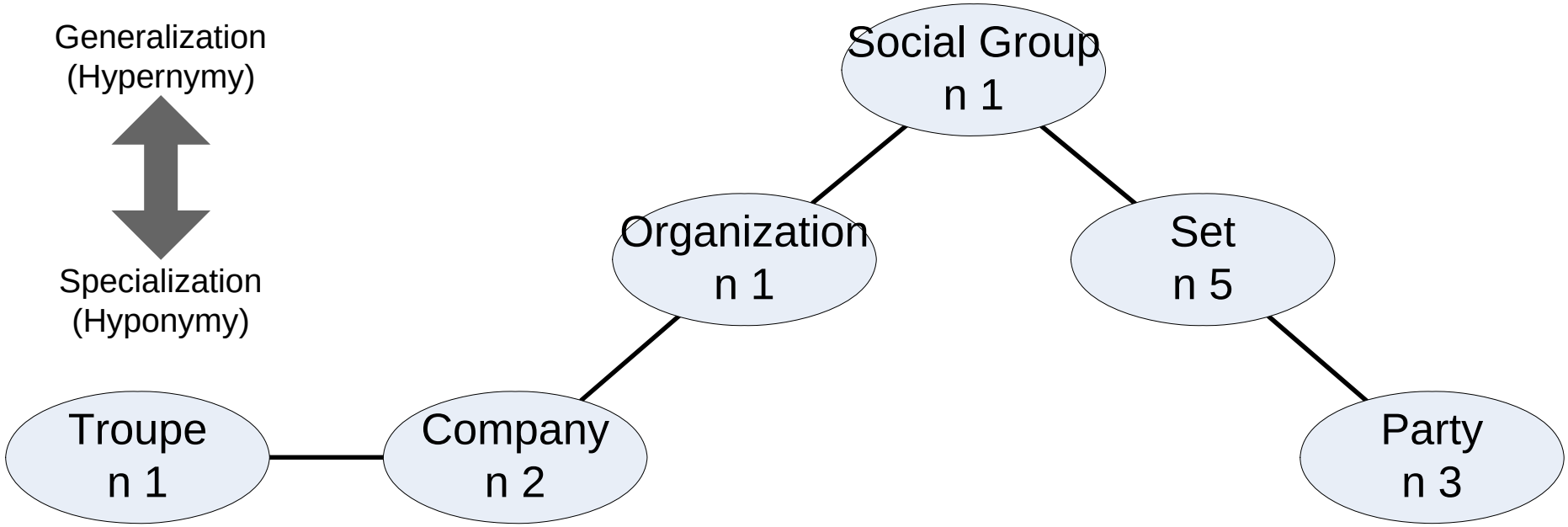
- **Build a graph, G, of (word, sense) pairs**
  - WordNet



PURDUE UNIVERSITY

Equimark: Embedding

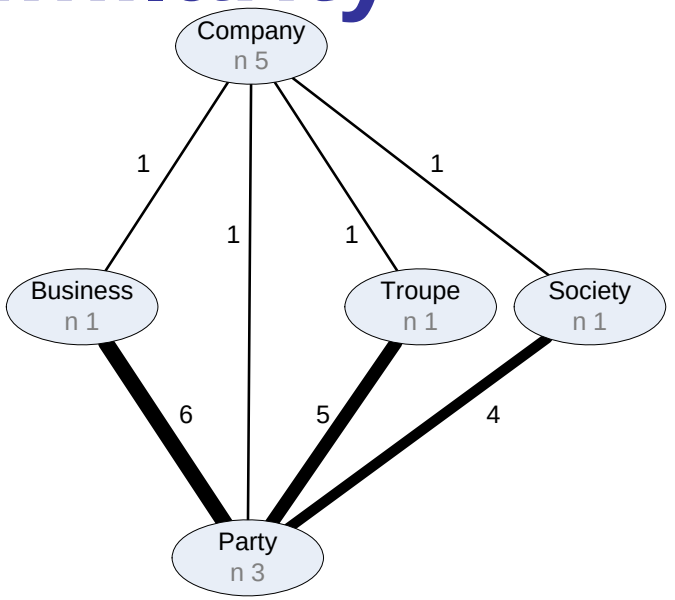
- **Build a graph, G, of (word, sense) pairs**
  - WordNet
- **Assign weights to the edges**
  - Using a “word similarity measure”



PURDUE UNIVERSITY

Word Similarity

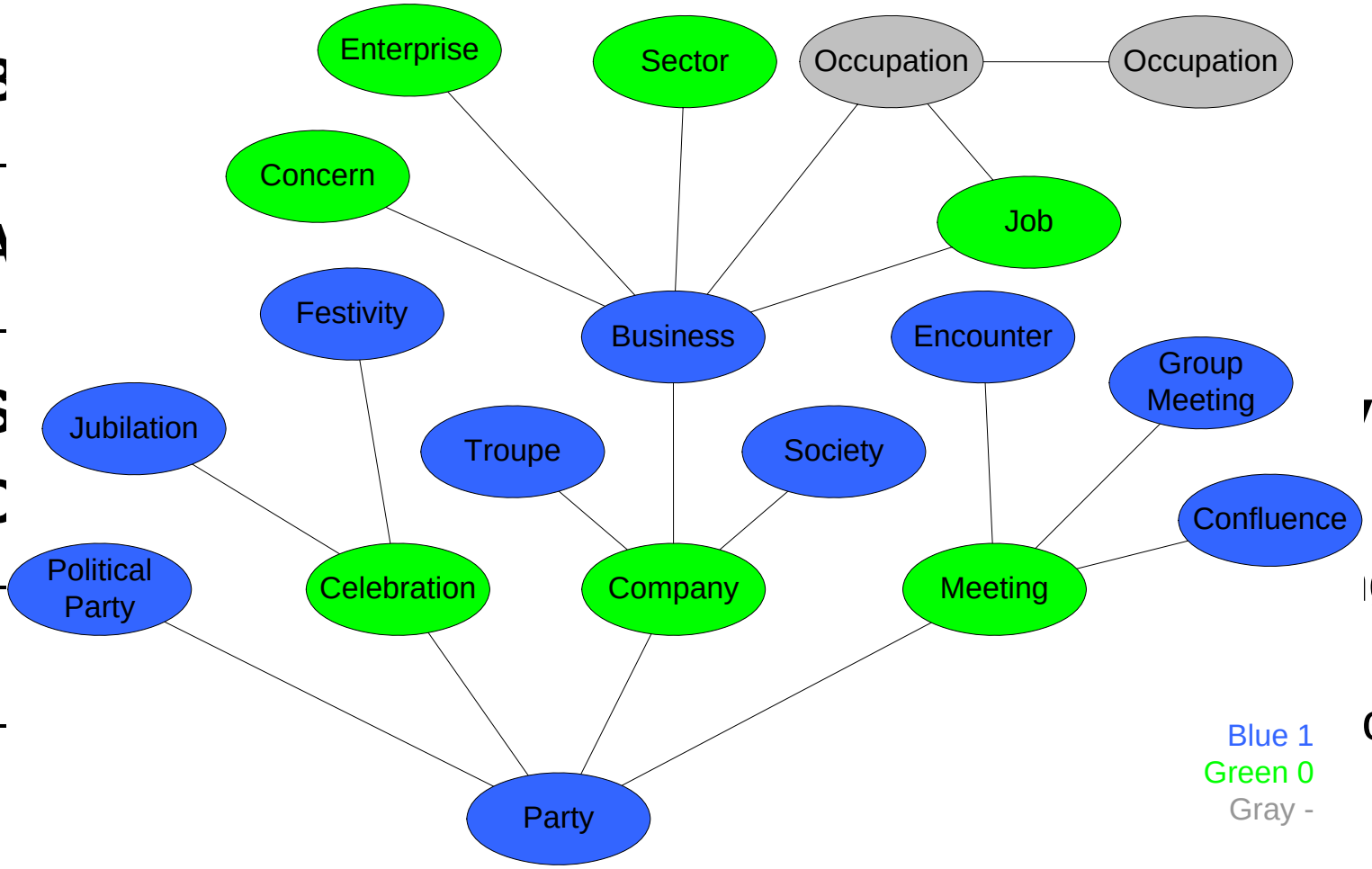
- **Not a Euclidean Space**



- **Word similarity fulfills other requirements expected from a distance function, $d_i()$**
  - Boundedness : finite distance between any given word pair
  - Symmetry: $d_i(a,b) = d_i(b,a)$
  - Equality: $d_i(a,b) = 0$ if and only if a = b

Equimark: Embedding

- **B**
  -
- **A**
  -
- **S** **', k**
- **C**
  - **g" to**
  -



ors

Quantifying Distortion

- **Watermark embedding distortion**

$$\sum_{s^N \in S^N} \sum_{k^N \in K^N} \sum_{m \in M} \frac{1}{|M|} p(s^N, k^N) d_1^N(s^N, f_N(s^N, m, k^N)) \le D_1$$

- **Maximum distortion an adversary can introduce**

$$\sum_{x^N \in X^N} \sum_{y^N \in Y^N} d_2^N(x^N, y^N) A^N(y^N \mid x^N) p(x^N) \le D_2$$

- **"Information-theoretic analysis of information hiding", P. Moulin and J. A. Sullivan, 2003**

PURDUE UNIVERSITY

Equimark: Embedding

- **for each word, w_i, in the given text**
  - bit_c = M[c]
  - if there is only one neighbor, w_c that encodes bit_c then replace **w_i**
  - if there are more than one neighbor that encodes bit_c
    - for each neighbor, w_c, of w_i calculate the expected distortion value for the adversary

$$E(d_2(w_c; w_i, s_k)) = \frac{\sum_{s_l \in S(w_c)} sim(w_c, s_l; w_i, s_k)}{|S(w_c)|}$$

    - pick the w_c that maximizes the $E(d_2(w_c; w_i, s_k))$

Equimark: Detection

- **Build the same graph, G, of (word, sense) pairs**
  - WordNet
- **Assign the same weights to the edges**
  - Using the same “word similarity measure”
- **Select the same sub-graph, G^W, of G using k**
- **Color G^W using k**
- **For each word in the watermarked text**
  - If color is gray skip
  - If color is blue concatenate 1 to M’
  - If color is green concatenate 0 to M’

Equimark in Action

They had to organize a party to search for help.

They had to form a company to seek aid.

They had to form a company to seek aid.

Make	Business	Get	Assistance
Organize	Troupe	Search	Help
Shape	Society	Attempt	Economic Aid
Forge	Party	Inquire	Care

They had to make a society to get assistance.

They had to forge a business to get economic aid.

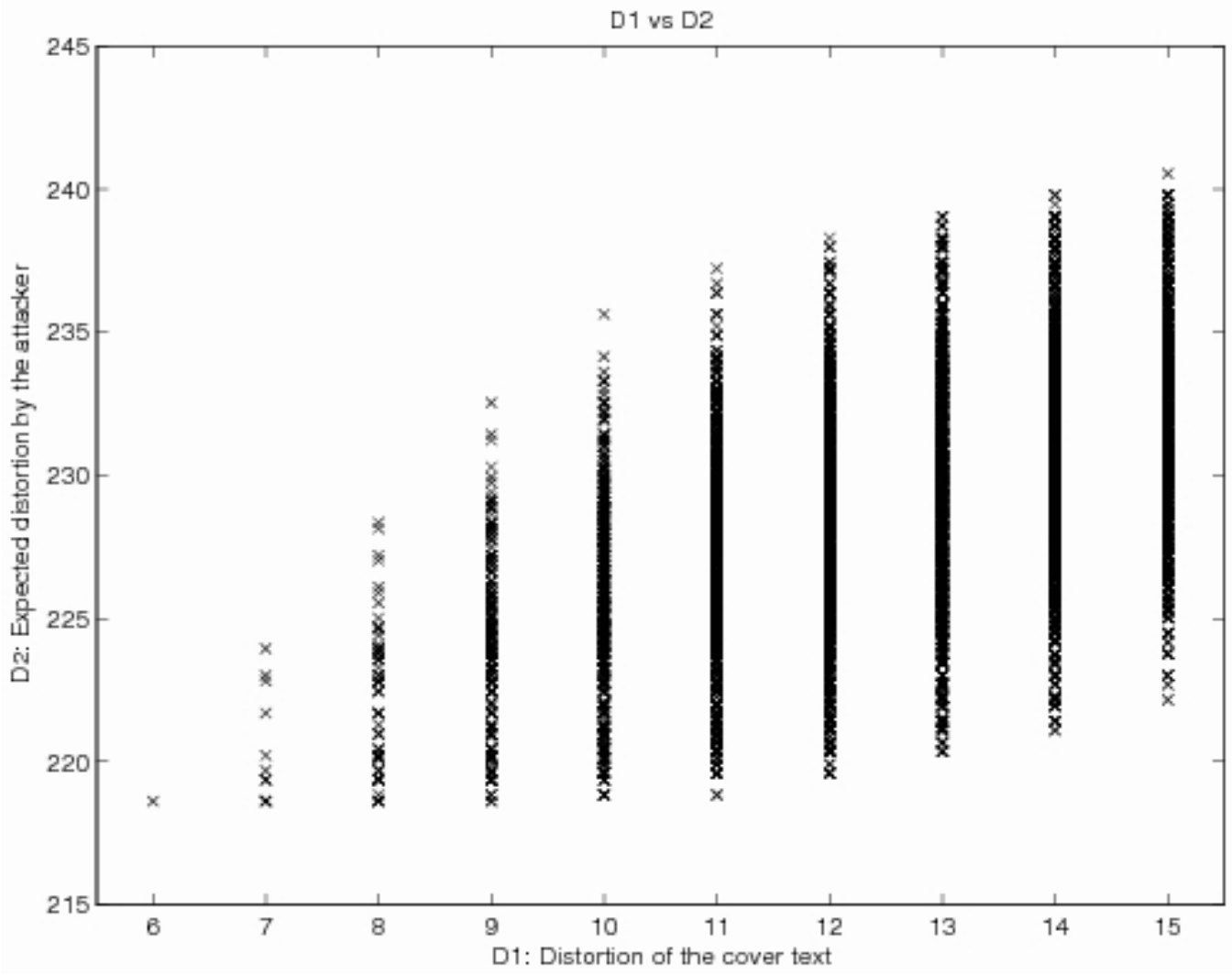
PURDUE UNIVERSITY

Building Equimark: Experiments

- **Data Resources:**
  - **A sense-tagged corpus**
    - **Semantic Concordance (Semcor 2.1)**
  - **WordNet**
    - http://wordnet.princeton.edu/perl/webwn
- **Software Resources:**
  - **WordNet::QueryData**
  - **WordNet::Similarity**
    - http://marimba.d.umn.edu/cgi-bin/similarity.cgi
    - We have used *pathlen()* as similarity metric

PURDUE UNIVERSITY

Quantifying Distortion



PURDUE UNIVERSITY

Conclusion

- **Protecting intellectual property rights for text**
  - How can you be sure that your articles/ papers/ blogs/ e-mails are not re-used?
  - Need a computationally light detection process (not AI complete)
  - Adversary can foil string matching
  - Even though you do not use watermarking, we can help
- **Embedding needs sense disambiguation, but detection**
- **Equimark embeds watermark into natural language text through *robust* synonym substitution**
  - achieves resilience by
    - giving preference to ambiguity-increasing transformations
    - using the maximum capacity below the distortion threshold

Future Work

- Enabling more domain-specific distance (word similarity) functions
- Increasing capacity and resiliency through the use of Wet Paper Codes and/or Error Correction Codes
- Testing copyright infringement detection performance
- Using a more powerful dictionary
  - "Bush returned to Washington D.C."
  - "The president came back to the capital"

Meaning Equivalent Changes

- **Context dependent synonyms**
  - **Semi automatic, interactive system**
  - **Adversary can not automate**

    “the sleuth” ⟷ “Sherlock Holmes” (A. C. Doyle)

    “the sleuth” ⟷ “Hercule Poirot” (A. Christie)
- **Generalization substitutions**
  - **moving up the “is_a” hierarchy**
  - **Adversary has to replace the general by the specific**

    “lion” ⟷ “big cat”

    “kangaroo” ⟷ “herbivore”

    “camel” ⟷ “herbivore”

    “kangaroo” ⟷ “marsupial”

PURDUE UNIVERSITY