CERIAS

the center for education and research in information assurance and security

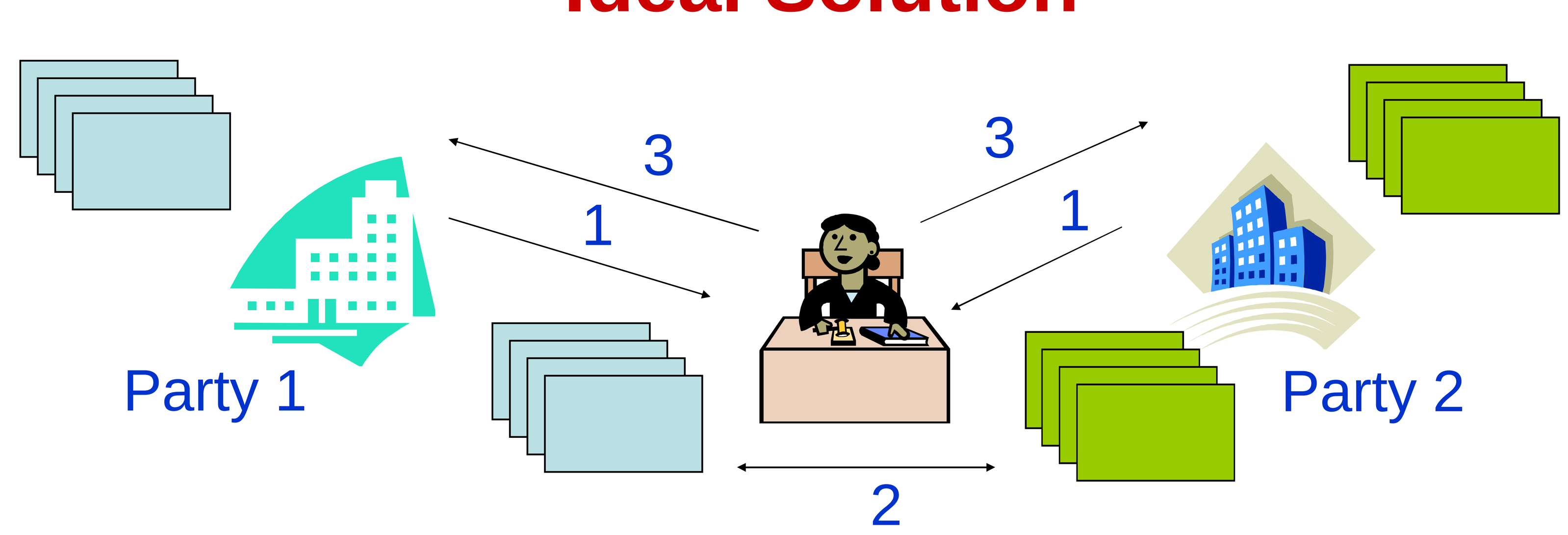
Secure Similar Document Detection

Wei Jiang, Mummoorthy Murugesan, Chris Clifton, Luo Si
Computer Sciences, Purdue University

Real World Scenario

- Two conferences need to find the double submissions, or plagiarized papers.
- Two companies want to find their common interests for a possible merger: e.g., # of similar or dissimilar projects
- Parties don't want to exchange the actual documents for several reasons:
  - Documents are confidential
  - Not to lose the competitive edge
  - Due to laws, and regulations

Ideal Solution



1. Two parties send their document collections to a Trusted Third Party (TTP).
2. The TTP performs the similarity test on the document collections
3. The TTP sends back only the similarity scores to the parties.

Do we really need a TTP? ***No***

Cosine Similarity Score

- Let Doc_1 = "develop drugs for AIDS, and TB"
  Doc_2 = "invent drugs for TB, and cancer"
  Term vector space={develop, invent, drugs, AIDS, TB, cancer}
- Documents are represented as vectors of term frequencies
  $D_1=(1,0,1,1,1,0)$
  $D_2=(0,1,1,0,1,1)$
- Normalize the vectors: $N_j = D_j/|D_j|$
  $|D_1|=2$, $|D_2|=2$
  $N_1=(.5,0,.5,.5,.5,0)$, $N_2=(0,.5,.5,0,.5,.5)$
- Cosine Similarity Score $(D_i,D_j) = D_i \bullet D_j / |D_i|\,|D_j| = N_i \bullet N_j$
  $sim(D_1,D_2) = 0.5$

Main Steps in SSDD

Step 1: Find Common Vector Space

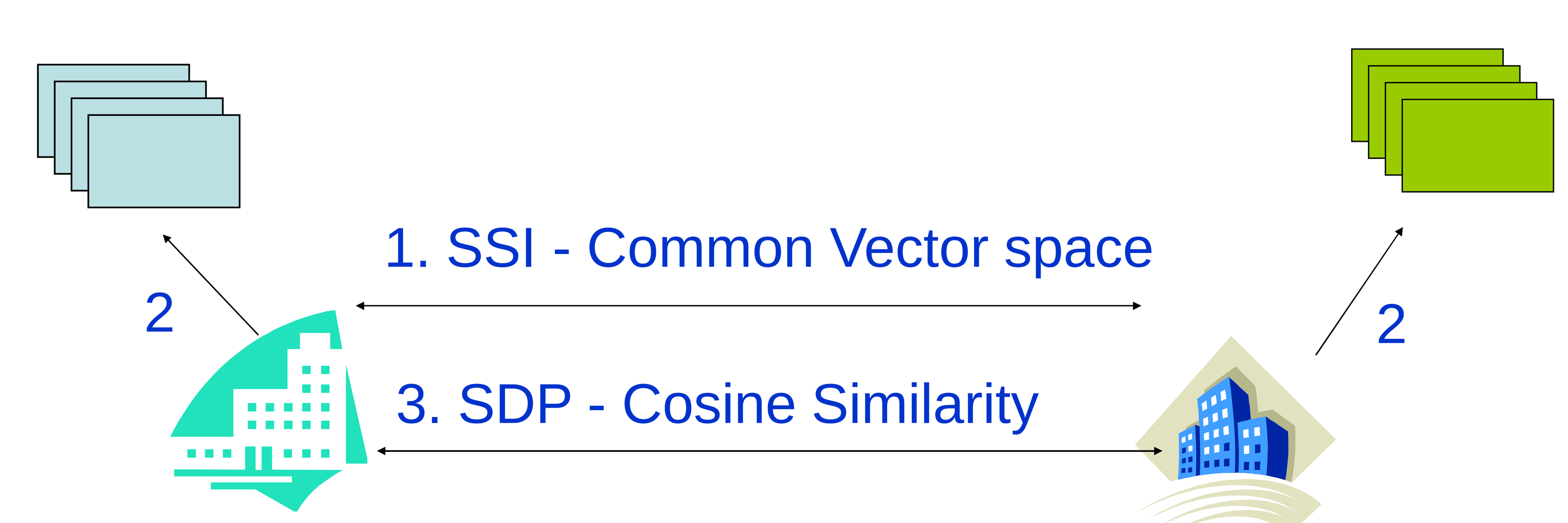
- Parties need to agree on a particular term vector space.
  $T_1 \cap T_2$ = {drugs, TB}
- Use a Secure Set Intersection (SSI) protocol to find the common vector space.
  Only $T_1 \cap T_2$ is revealed, neither T_1 nor T_2.

Step 2: Create Normalized vectors

- Each party computes the normalized term vectors of documents in their collections for $T_1 \cap T_2$
  $N_1' = (.5,0.5)$, $N_2' = (0.5,0.5)$

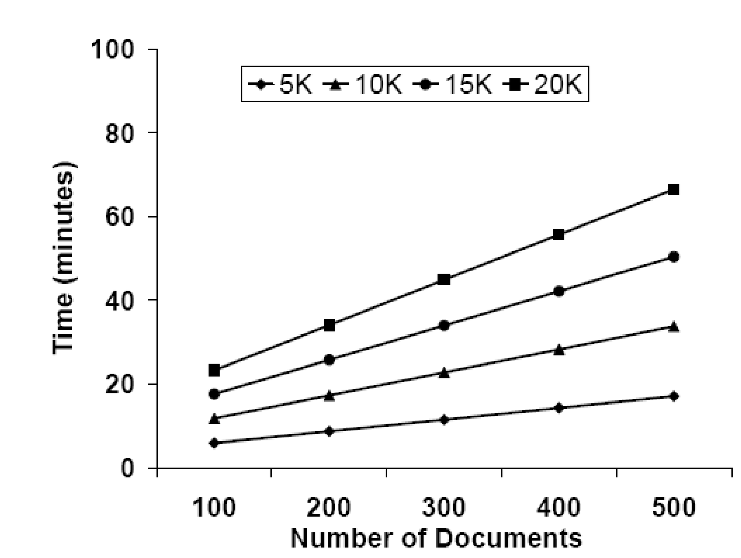
Step 3: Compute Dot Product Securely

- Use any one type of the Secure Dot Product (SDP) protocol:
  1) Random Matrix (Vaidya & Clifton, 2002)
  2) Homomorphic encryption scheme (Goethals, et al., 2004)
- Compute pair-wise document cosine similarity using the SDP protocol on the normalized document vectors.
  $N_1' \bullet N_2' = 0.5$ (similarity score between Doc_1 and Doc_2)
- Actual values in the normalized vectors are not revealed.



Experiments

- Tested on a document collection consisting of 100 papers from a major DB conference (# of terms is ~15k)
- Tested on randomly generated vectors
  - The plot shows the running time of SSDD_H
  - Number of documents vary from 100 to 500
  - Number of terms vary from 5K to 20K



Conclusion

- Experimental results show that SSDD is practical for reasonably large collection of documents
- Can be used as a preliminary step to get candidates for more manual/comprehensive checks

PURDUE UNIVERSITY

CERIAS

Discovery Park
e-Enterprise Center