SAND2017-9546C

Compression algorithms for deception detection

SOCINFO 2017 CONFERENCE, OXFORD

CHRISTINA L. TING

ANDREW N. FISHER

TRAVIS L. BAUER

SANDIA NATIONAL LABORATORIES

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. De- partment of Energy's National Nuclear Security Administration under contract DE-NA0003525

Motivation: Deception Detection

Deception is a common element in many forms of communication

- Analysts need tools for processing huge amounts of data
- Human judgement performs no better than chance

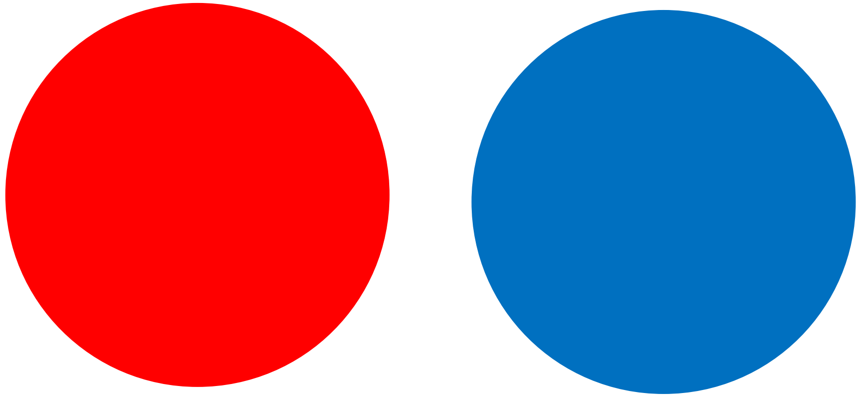
Current methods for deception detection based on supervised learning

- Requires manual feature selection
- Based on psycholinguistic theories of deception
- And/or computational linguistics

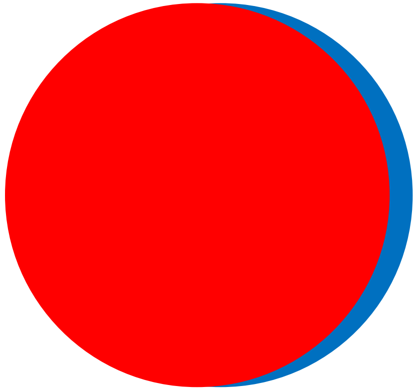
Compression as a featureless method

- Compression algorithms aim to reduce a document’s size
- Effectively, to use fewer bits for the same information
- Reduction in size is possible due to
  - Information redundancy
  - Information similarity
  - Predictable structure
- Idea: Use these properties to identify similarities in documents.

Clustering based on NCD



NCD = 1



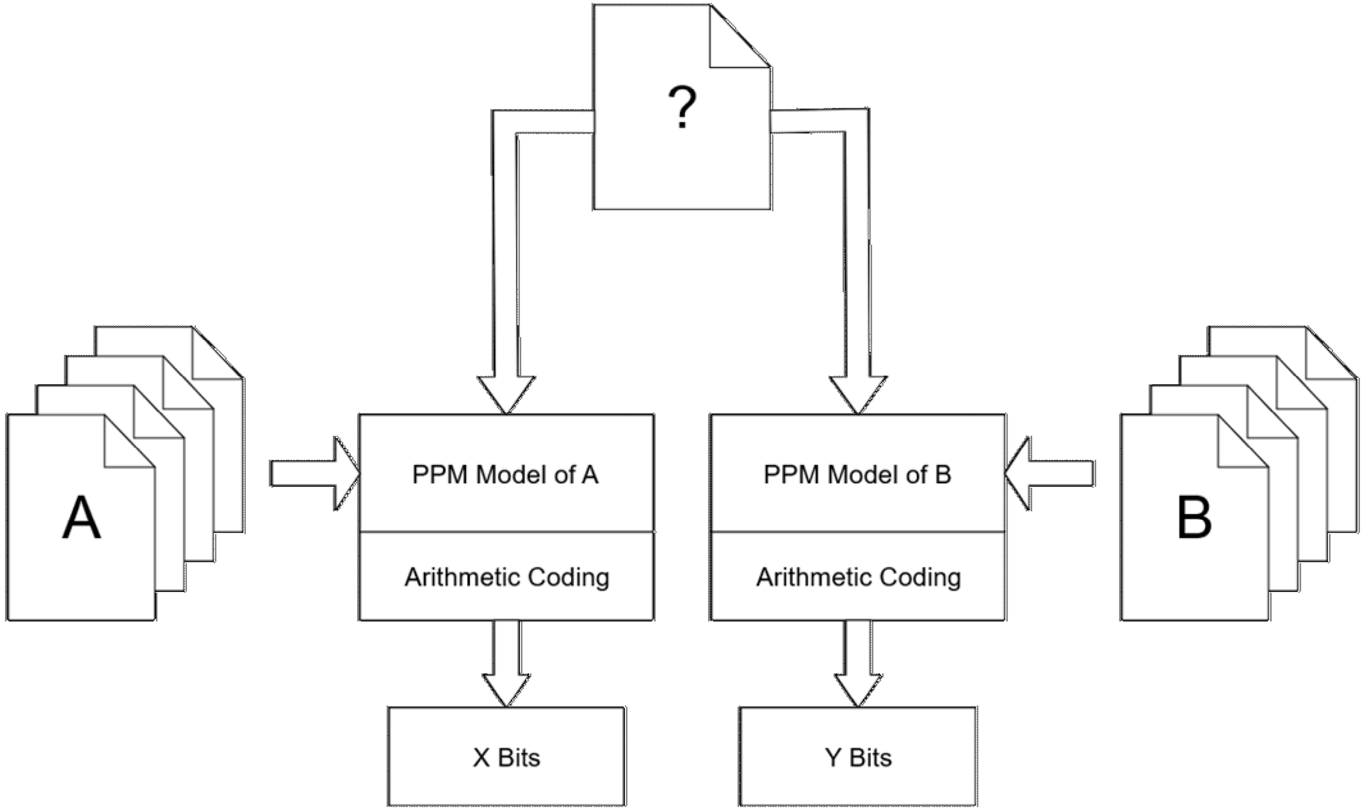
NCD ~ 0

$$NCD(x,y) = \frac{C(xy) - \min(C(x), C(y))}{\max(C(x), C(y))}$$

C(x) is a compression algorithm applied to item x.

We use LZMA.

Li, Chen, Li, Ma, and Vitanyi, The Similarity Metric, IEEE Transactions on Information Theory, 2004



Classifying based on PPM

PPM is based on Markov models of different orders.

A character is predicted based on the previous *n* characters.

Context *P(c|predi)*

Brennan-Greenstadt (BG) deception dataset

1. B-G corpus (12 authors):
   - Truthful texts: ~5000 + words of pre-existing author samples
   - Obfuscation text: ~500 word description of neighborhood while trying to hide identity
   - Imitation text: ~500 word article in the style of Cormac McCarthy
   - In total: 24 deceptive texts (12 each of obfuscation and imitation), 113 truthful texts
2. Extended B-G corpus:
   - Similar to above, except 45 author submissions from AMT

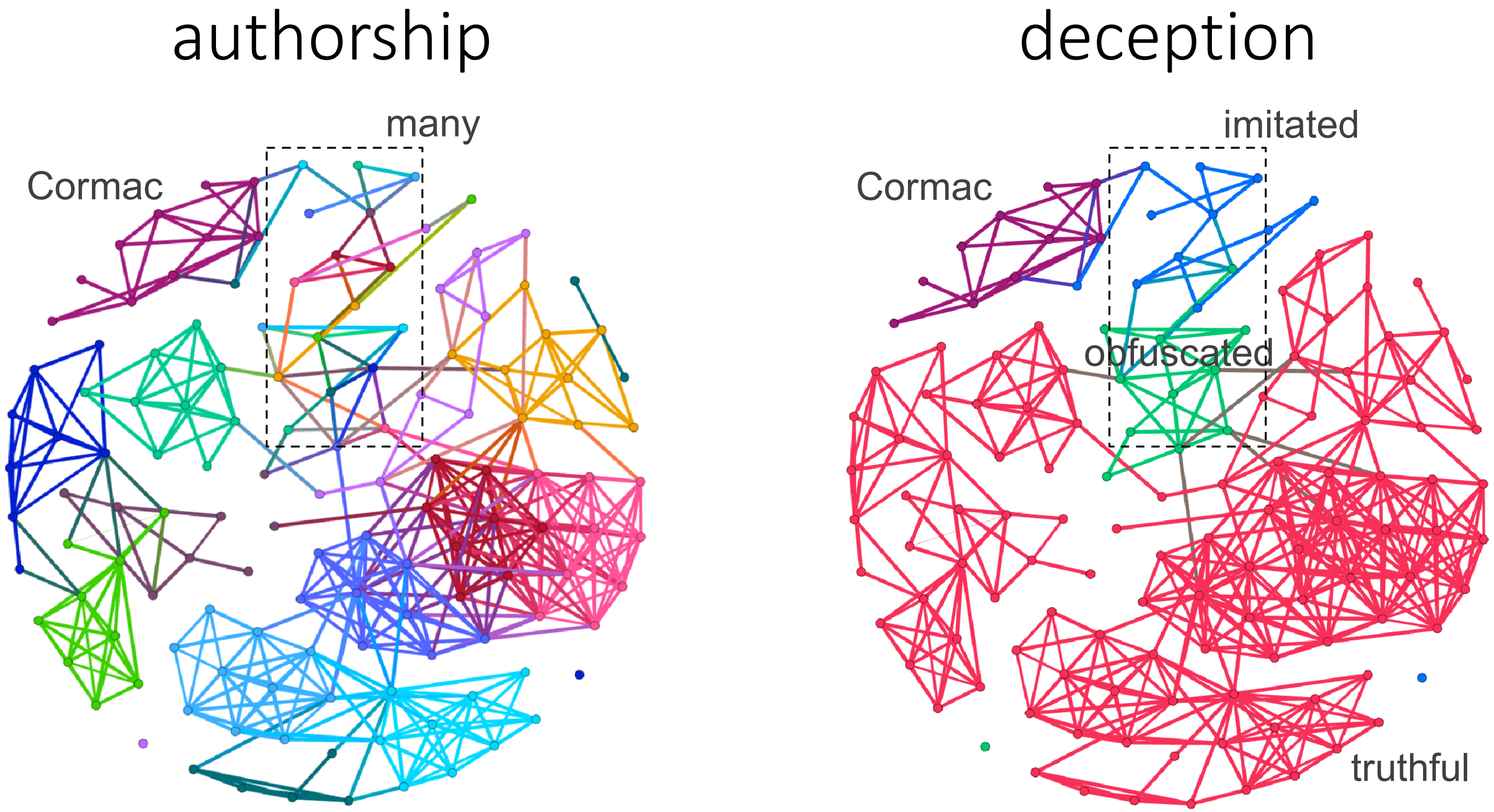
[2] https://psal.cs.drexel.edu/index.php/Main_Page

Who’s the real Cormac?

“Leaning with his head practically touching his knees, he sat at the edge of the bed, waiting for consciousness to take hold. He waited for his eyes to adjust to the soft, pale light and stood. Hoping to find a note he checked his desk for any stray scraps of paper. Nothing.”

“With the first gray light he rose and left the boy sleeping and walked out to the road and squatted and studied the country to the south. Barren, silent, godless. He thought the month was October but he wasnt sure. He hadnt kept a calendar for years. They were moving south.”

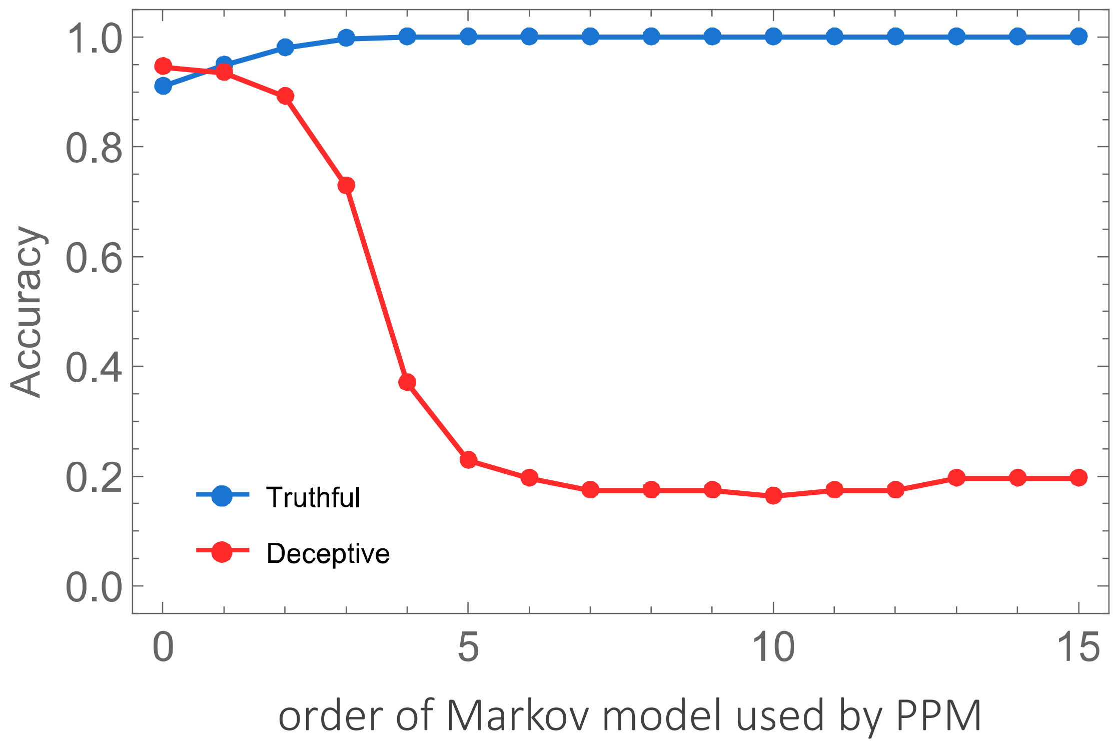
Results: NCD clustering

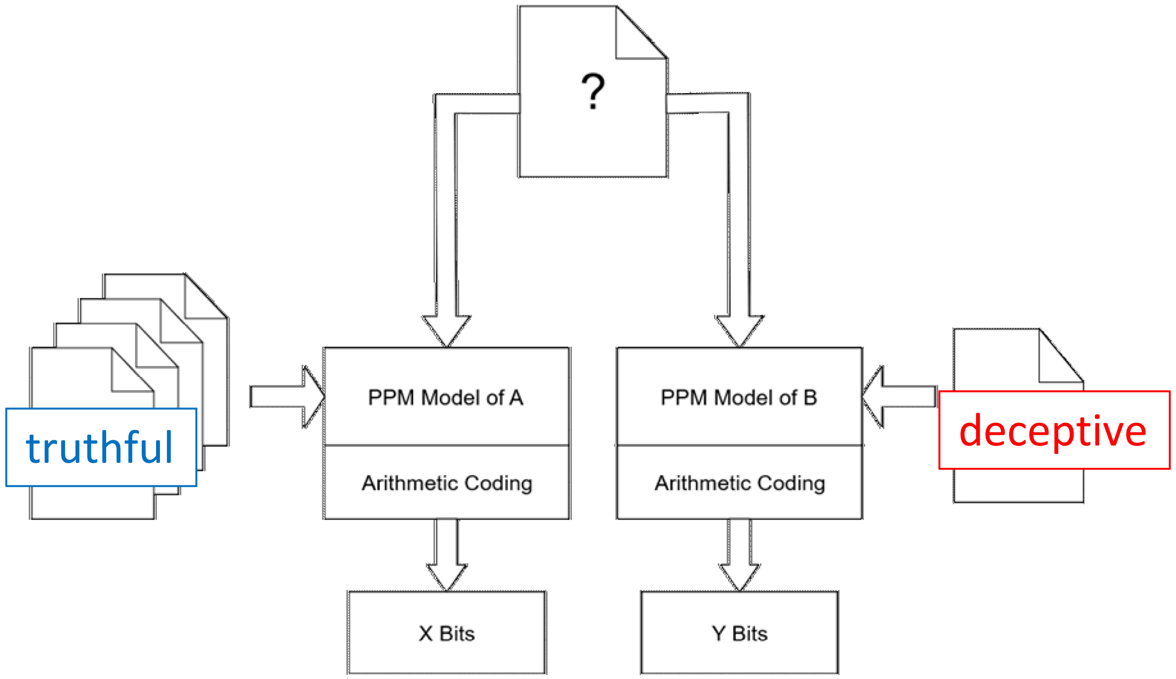


- Successfully applied to authorship attribution.
- Identifies deception across several authors.
- Easy to hide one's identity.
- Difficult to imitate another's identity.

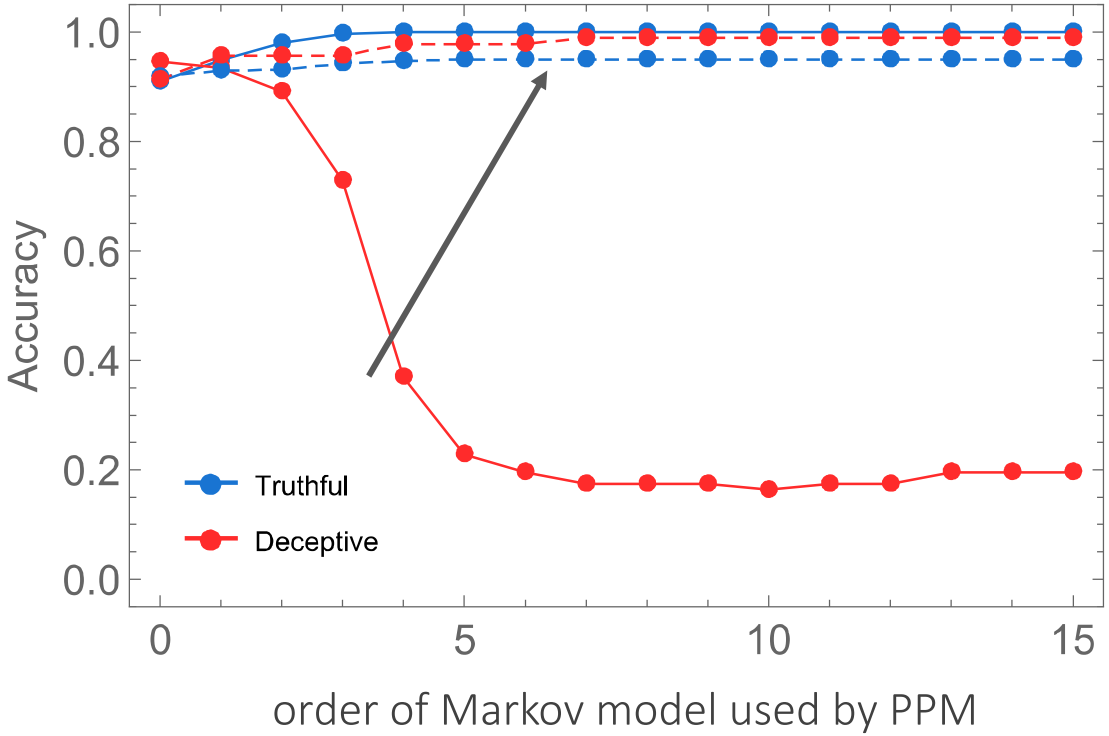
[2] https://psal.cs.drexel.edu/index.php/Main_Page

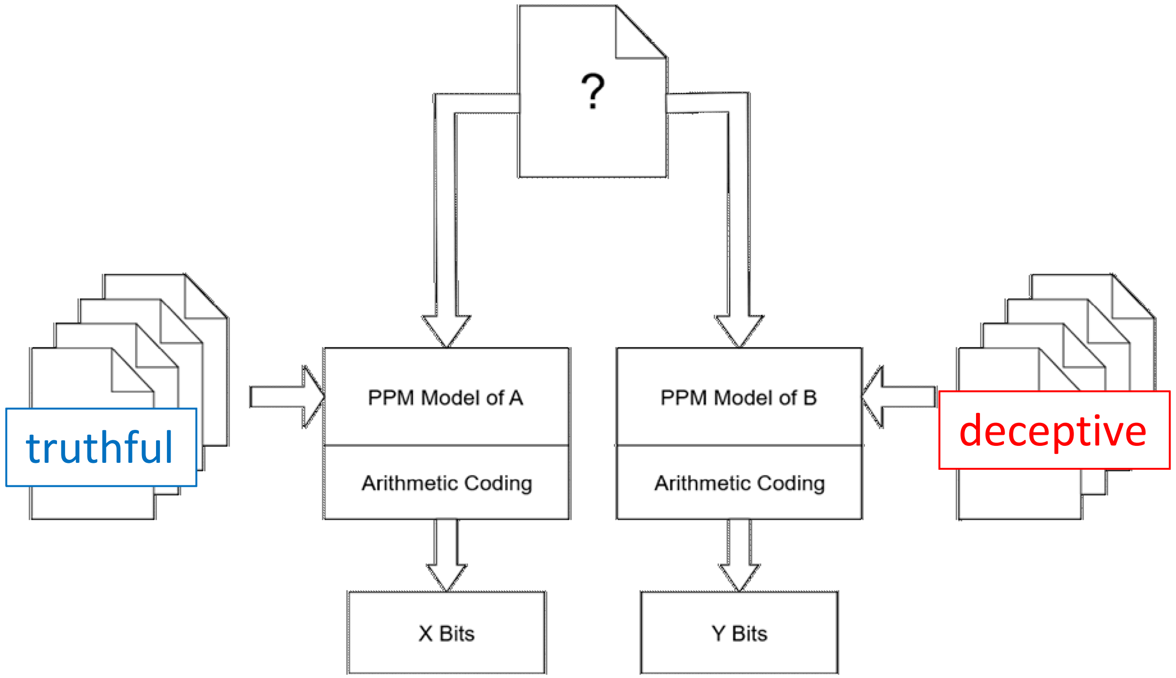
Results: PPM classification





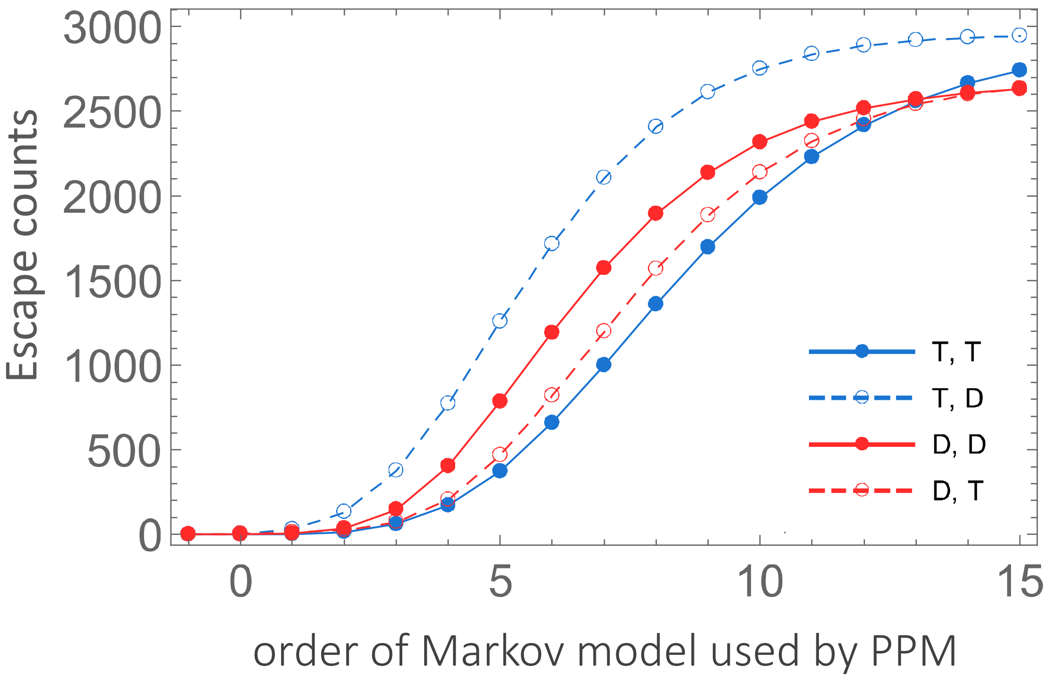
First solution: balance the models



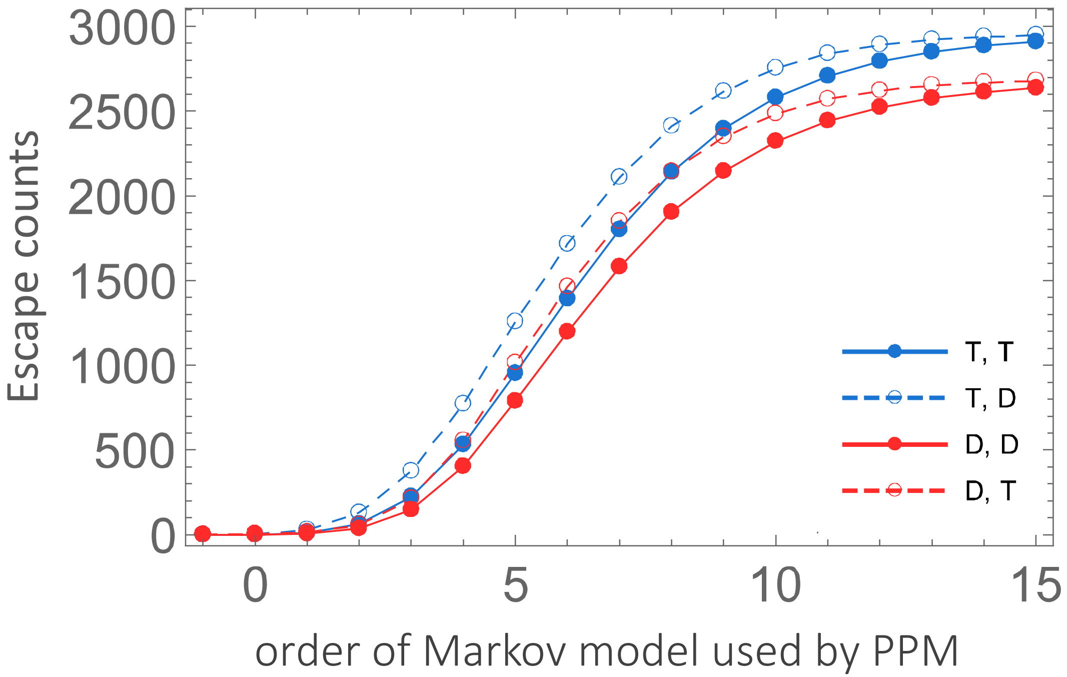


A closer look at the escapes

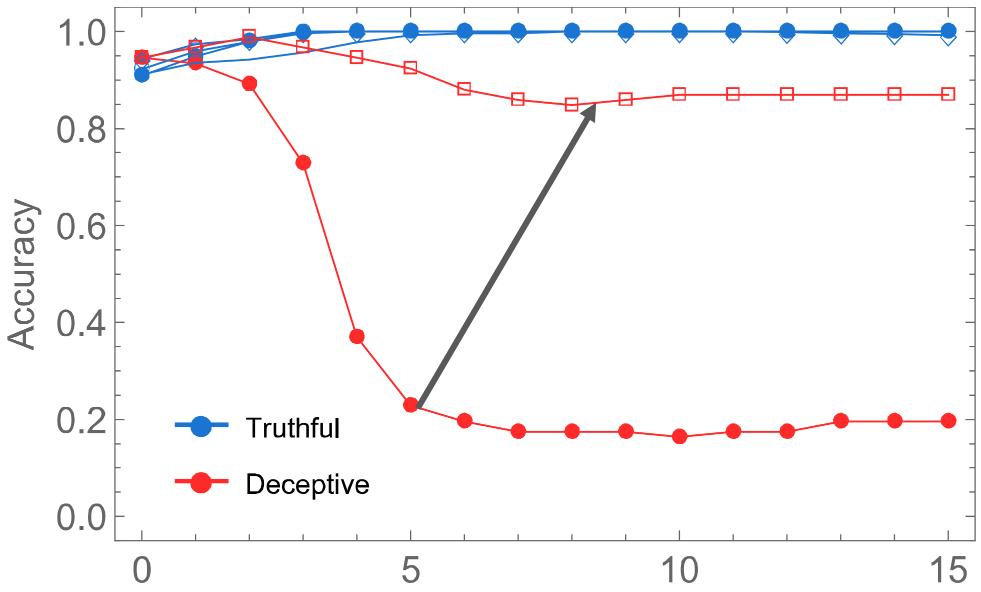
Unbalanced



Balanced



Second solution: remove the escapes



order of Markov model used by PPM

- Escapes signal to the decompressor that a lower context was used
- We don’t need the escapes for analytics; drop them

Conclusions

- Compression algorithms successfully implemented in featureless clustering and classification techniques.
- Modifications to the algorithms can be made to enhance their use in analytic techniques.
- Future work:
  - Determine what structure compression identifies.
  - Modify compression algorithm to identify that structure.

Questions?