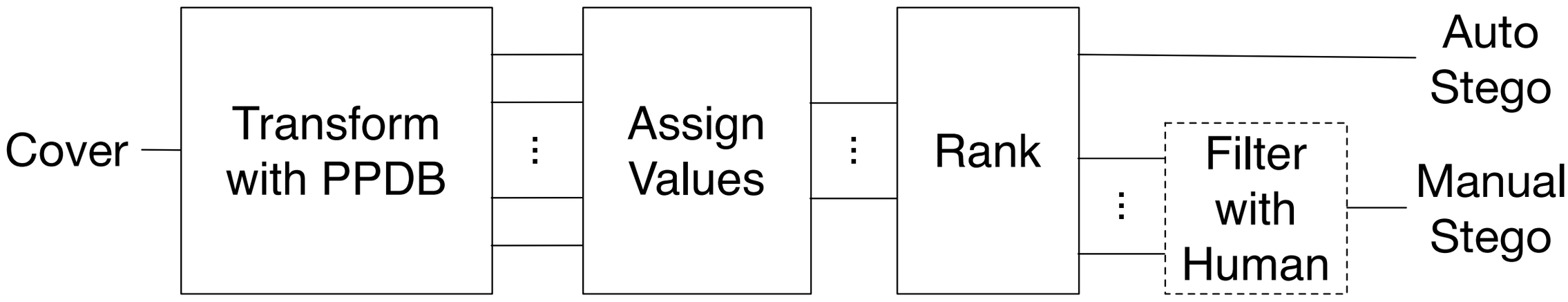
Distortion Measures for Linguistic Steganography

Alex Wilson, Andrew Ker
University of Oxford
16th Feb, Electronic Imaging 2016

The CoverTweet System



Attack

- Follows standard paradigm of steganalysis: large feature set used to train a classifier.

- Four classes of features: basic, probability, word length, PPDB.

Features

1. Basic.

    Mean and variance of word length, number of words, individual stop word counts.

I just love pizza

Features

2. Probability.

    Mean and variance of n-gram probabilities, for n from 1 to 5.

Pr(‘pizza’ | ‘I’, ‘just’, ‘love’)

Features

3. Word length.

    Mean and variance of word length sequence probabilities, for sequence length 1 to 10.

Pr(‘pizza’ | ‘I’, ‘just’, ‘love’)

Pr(5 | 1, 4, 4)

Features

4. PPDB.

Features extracted using CoverTweet's transformation source, including the probability of the most likely transformed sentence.

$$-\log(\max_{t \in \mathbf{T}} \Pr(t))$$

Pooled Steganalysis

203 features

thank god i woke up
ahhh i need a ride to school ...
definitely should have gone to school...
lol
doing nothing...
i don't look like im 70... i barely look like im 18...
right ? ! sad day...
am i right though?!
-___-
soooo attractive! ugh!
i don't like thinking about this.

812

Value Assignment

Could number substitutions alphabetically:

0	progressed
1	strolled
2	travelled
3	wandered
4	walked

0	bank
1	river
2	riverbank

She walked[4] to the bank[0]

Value Assignment

Doesn't work if the sets are non-disjoint:

0	progressed
1	strolled
2	travelled
3	wandered
4	walked

0	bank
1	river
2	riverbank

0	bank
1	store
2	treasury
3	fund

She walked[4] to the bank[?]

Value Assignment

Alternatively, could just use a hash function:

01 She strolled to the river
10 She wandered to the store
11 she walked to the bank
00 She travelled to the riverbank

Value Assignment

I hate geometry

I just hate geometry

I detest geometry

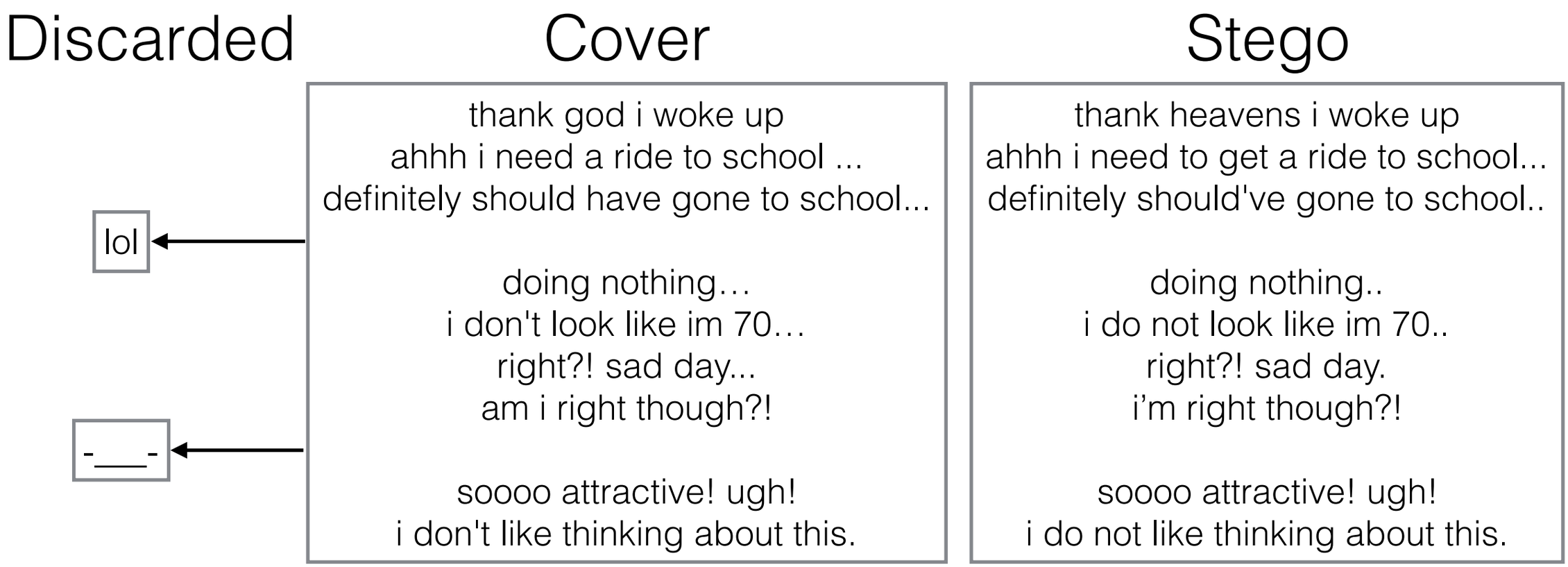
I loathe geometry

I am geometry

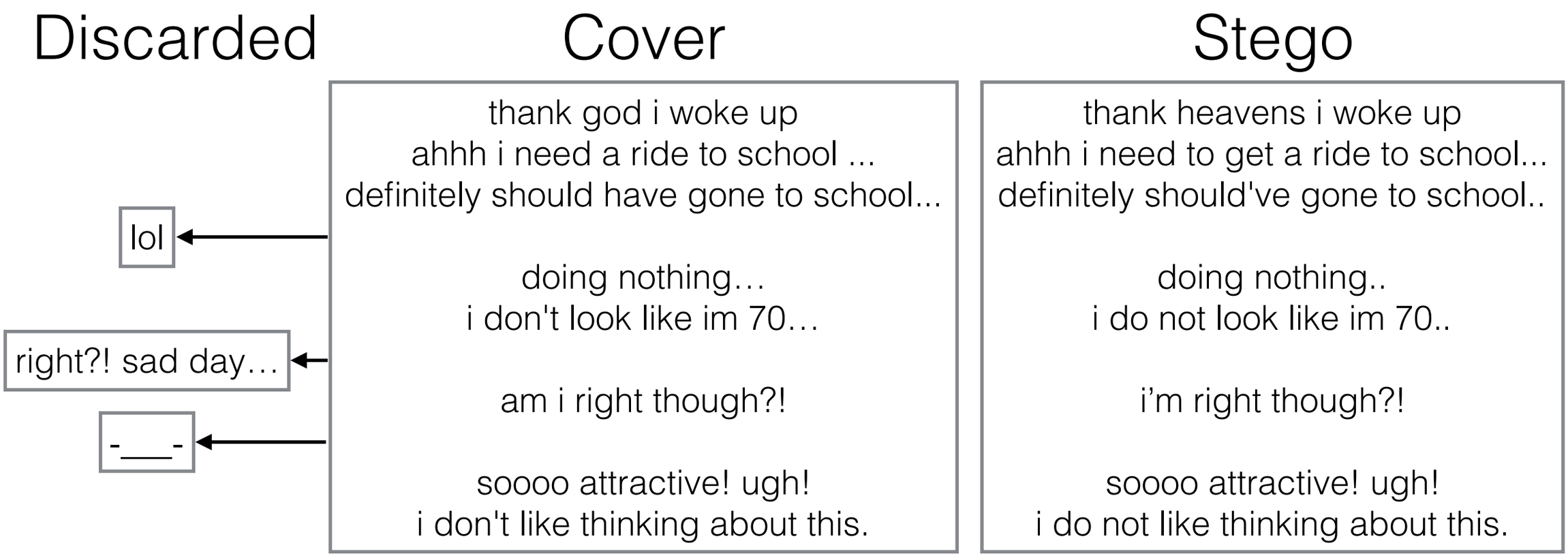
Value Assignment

000 I loathe geometry
001
010 I detest geometry
011
100 I just hate geometry
101 I am geometry
110
111 **I hate geometry**

Data Generation



Data Generation



Source Coding

We can spread the payload out across multiple tweets, using *source coding*.

- Solves the selection channel problem.
- Improves efficiency.
- Minimises total distortion.

Distortion

000	I loathe geometry	3
001		inf
010	I detest geometry	2.6
011		inf
100	I just hate geometry	1
101	I am geometry	10
110		inf
111	**I hate geometry**	0

Distortion

111	**I hate geometry**	0
100	I just hate geometry	1
010	I detest geometry	2.6
000	I loathe geometry	3
101	I am geometry	50
001		inf
011		inf
110		inf

Source Coding

- The cost of embedding symbol *j* in cover *i* is d_{ij}.
- The probability of this change being made is p_{ij}.

$$p_{ij} = \frac{\mathrm{e}^{-\lambda d_{ij}}}{\sum_j \mathrm{e}^{-\lambda d_{ij}}}$$

We *simulate* perfect coding by finding the λ that gives us the desired payload size *m*.

Distortion Measures

1. Binary.
2. Probability.
3. Edit distance.
4. Feature.

Distortion Measures

1. Binary

Simple binary distortion:
Unchanged = 0
Changed = 1

Distortion Measures

2. Probability

Log likelihood ratio between the cover and the stego:

$$-(log\,\Pr(s) - log\,\Pr(c))$$

Distortion Measures

3. Edit distance

Minimum number of word insertions, deletions and substitutions required to turn the cover into the stego.

I hate geometry

↓

I just don't like geometry

(Edit distance 3)

Distortion Measures

4. Feature distortion

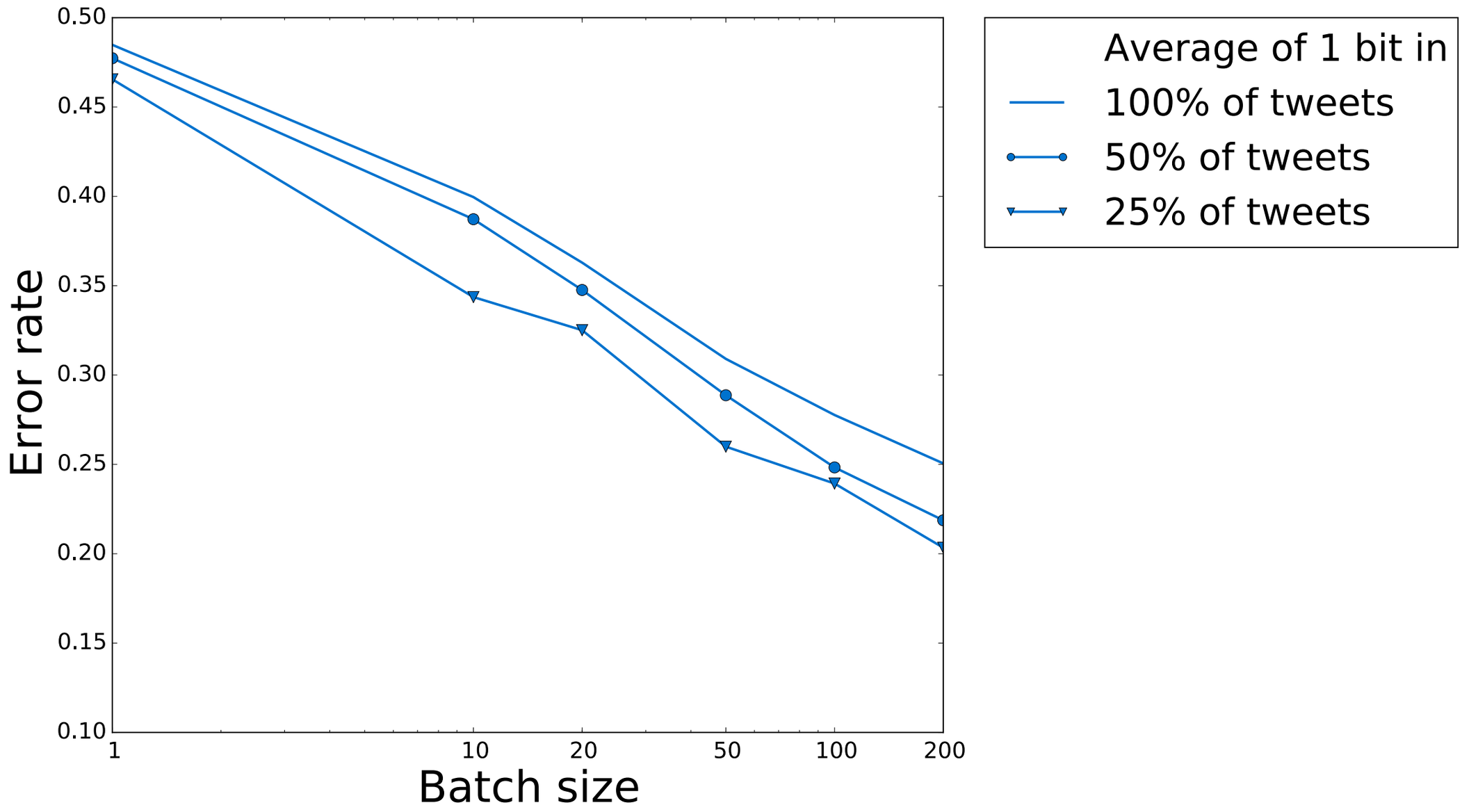
The distance between the feature vectors for the cover and the stego. Requires knowledge of the steganalysis features.

Testing data

- Automatic stego: 500 users, 1000 tweets each. Average of 1 bit coded in 100%, 50% and 25% of tweets.
- Manual stego: 10 users, 100 tweets each. Average of 1 and 2 bits coded in 100% of tweets.
- Features pooled with a range of batch size.

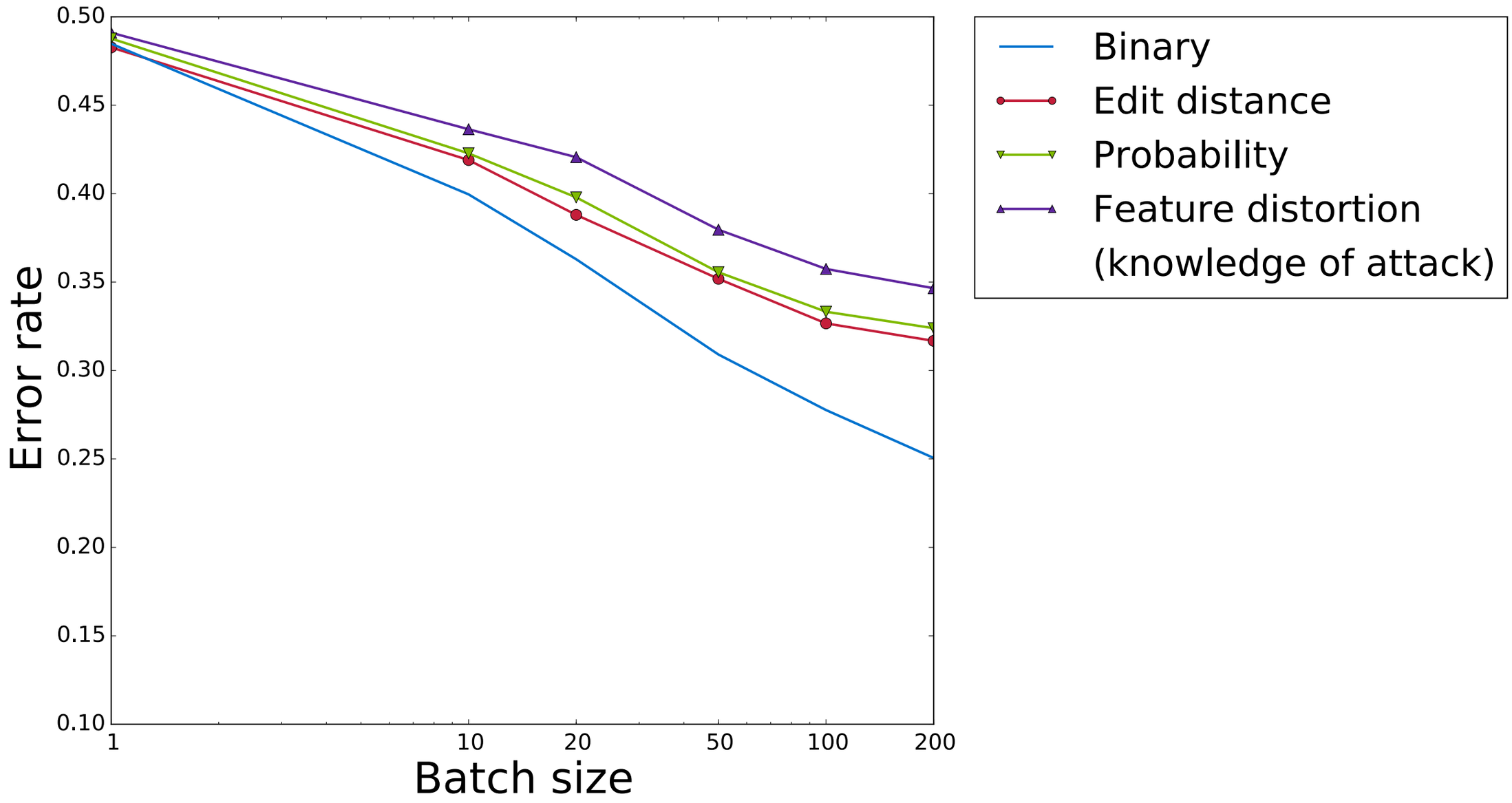
Automatic Results

Binary distortion:



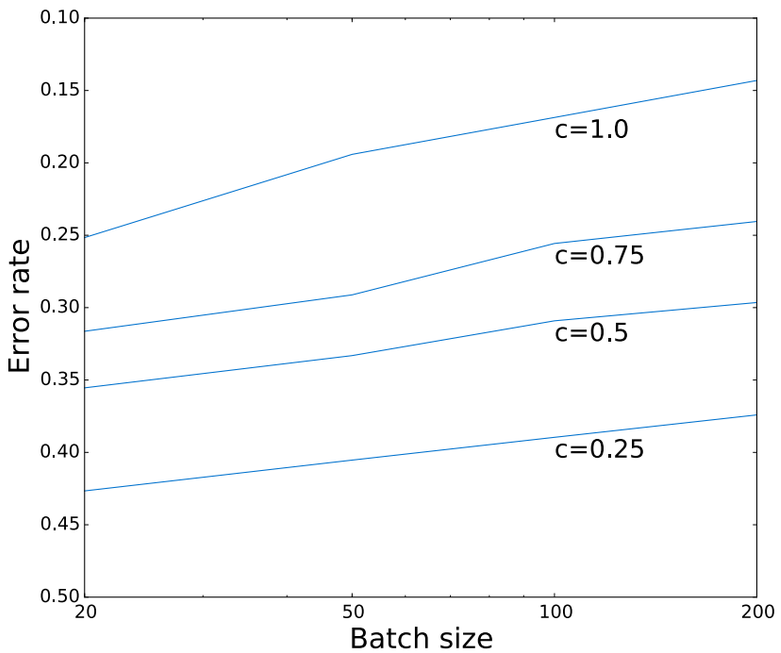
Automatic Results

All distortions:

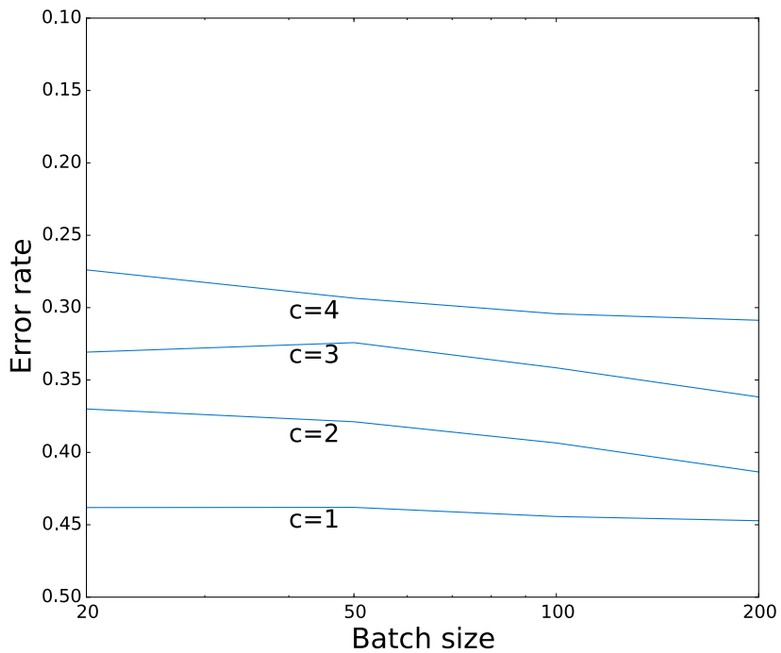


Square root law

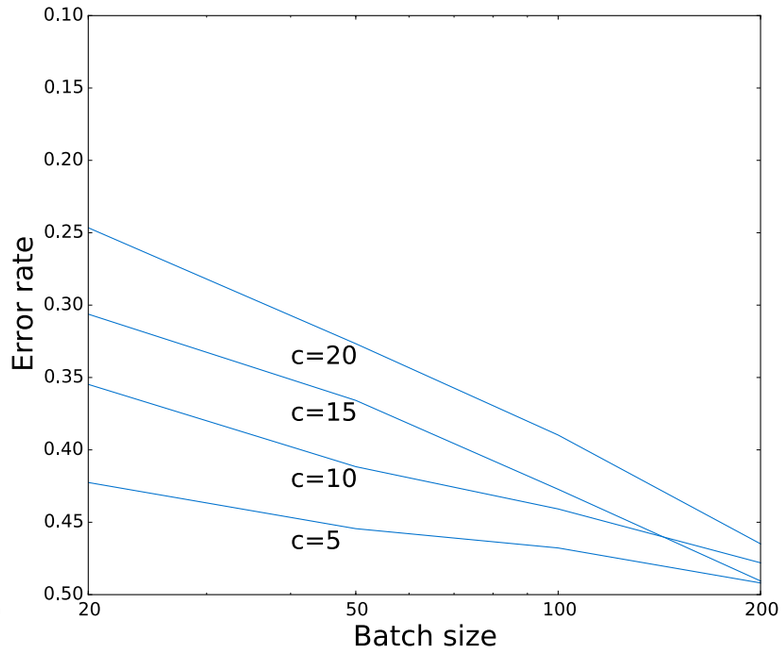
payload prop. to # covers



payload prop. to sqrt. # covers

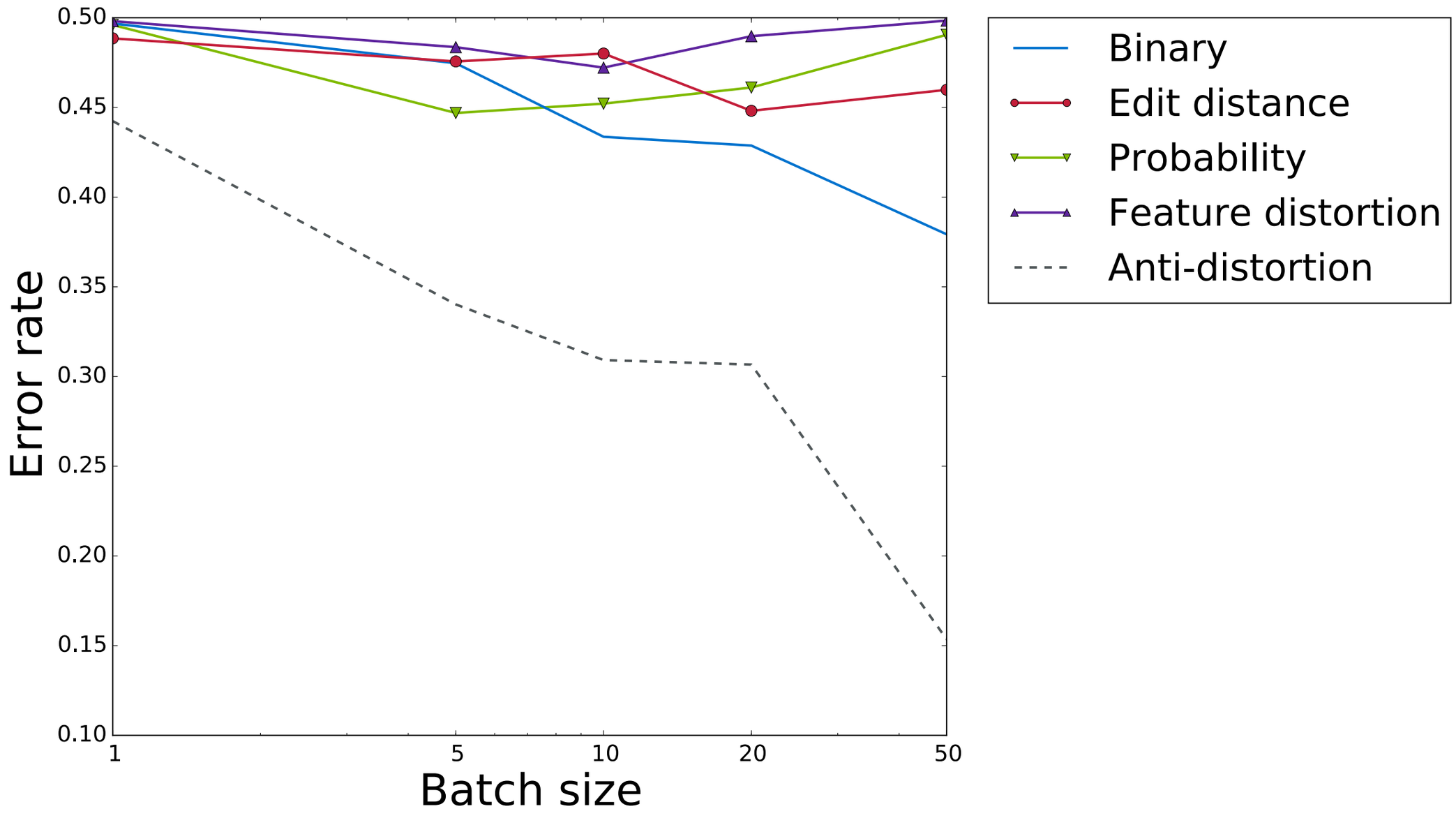


payload constant



Manual Results

All distortions:



Summary

- Coding is crucial for linguistic steganography.
- We have introduced the first distortion measures for linguistic steganography.
- If we ever want to detect manually filtered stego, we need a lot more data.
- The square root law is apparent in linguistic steganography.