Linguistic Steganography on Twitter: Hierarchical Language Modelling with Manual Interaction

Alex Wilson, Phil Blunsom, and Andrew D. Ker
University of Oxford

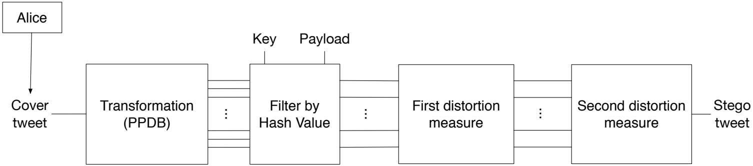
Twitter

- Twitter is a social networking site, launched in 2006.
- Users post short messages (*tweets*), at most 140 characters long.
- 500M tweets posted each day, from 200M active users.
- Twitter a suitable setting because linguistic steganography generally requires the steganographer to act as the cover source.

Twitter Steganography

- Alice has a Twitter account, and has posted some number of innocent tweets, before starting to send steganographic messages.
- Bob shares a key with Alice, and has access to her tweets.
- We assume the Warden is human.

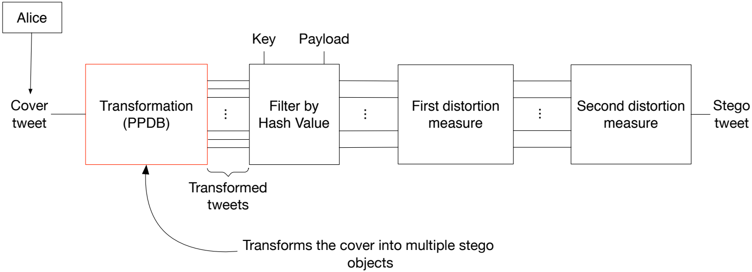
CoverTweet



Alice
Key
Payload
Cover tweet
Transformation (PPDB)
Filter by Hash Value
First distortion measure
Second distortion measure
Stego tweet

CoverTweet

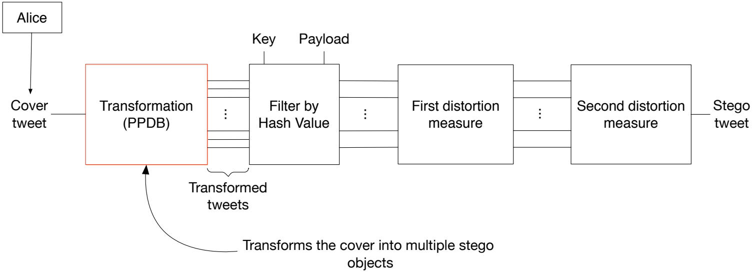
```
gosh now I really don’t want my beard to go away
```



CoverTweet

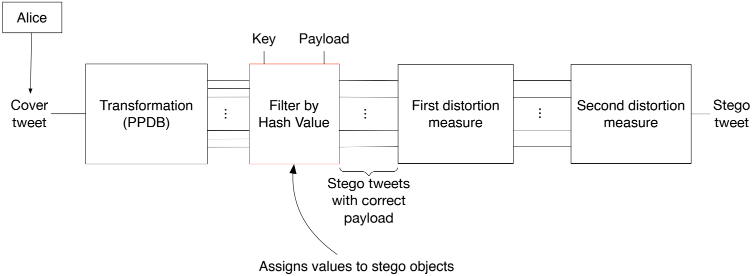
gosh now I really don’t want my beard to go away

gosh today i truly don’t want anything my beard to move away
gosh now i genuinely don’t want my beard to go away
god now i truly do not want my beard to go away
gosh today i really don’t want my beard to go away
...
gosh now I really don’t want to my barbe of going away
gosh now I genuinely just don’t wanna my beard to go away
gosh there, i really don’t wanna my beard to go away
gosh now I really don’t mean my beard to get away
gosh now I truly don’t want my beard of going away



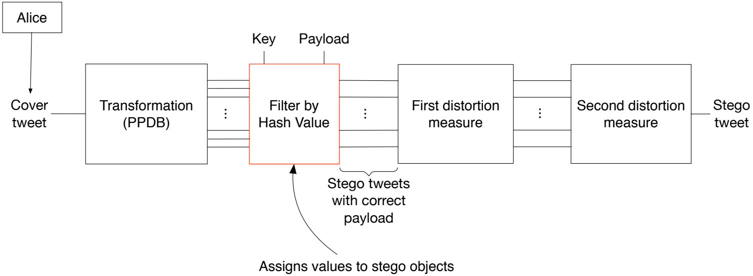
CoverTweet

```
gosh today i truly don't want anything my beard to move away
          gosh now i genuinely don't want my beard to go away
              god now i truly do not want my beard to go away
           gosh today i really don't want my beard to go away
                                                          ...
       gosh now I really don't want to my barbe of going away
    gosh now I genuinely just don't wanna my beard to go away
         gosh there, i really don't wanna my beard to go away
            gosh now I really don't mean my beard to get away
          gosh now I truly don't want my beard of going away
```



CoverTweet

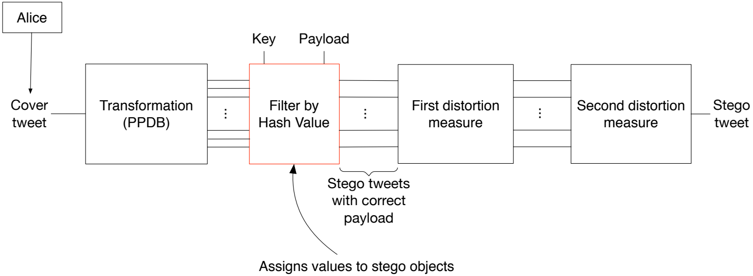
gosh today i truly don't want anything my beard to move away	0100
gosh now i genuinely don't want my beard to go away	0100
god now i truly do not want my beard to go away	1100
gosh today i really don't want my beard to go away	0110
...	
gosh now I really don't want to my barbe of going away	0001
gosh now I genuinely just don't wanna my beard to go away	0100
gosh there, i really don't wanna my beard to go away	1101
gosh now I really don't mean my beard to get away	0110
gosh now I truly don't want my beard of going away	0100



CoverTweet

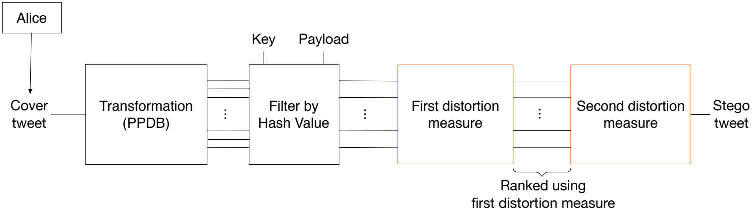
```
gosh today i truly don't want anything my beard to move away 0100
        gosh now i genuinely don't want my beard to go away 0100
            god now i truly do not want my beard to go away 1100
         gosh today i really don't want my beard to go away 0110
                                                        ...
     gosh now I really don't want to my barbe of going away 0001
  gosh now I genuinely just don't wanna my beard to go away 0100
       gosh there, i really don't wanna my beard to go away 1101
          gosh now I really don't mean my beard to get away 0110
          gosh now I truly don't want my beard of going away 0100
```

```
0100 gosh today i truly don't want anything my beard to move away
0100 gosh now i genuinely don't want my beard to go away
0100 gosh now I genuinely just don't wanna my beard to go away
0100 gosh now I truly don't want my beard of going away
```



CoverTweet

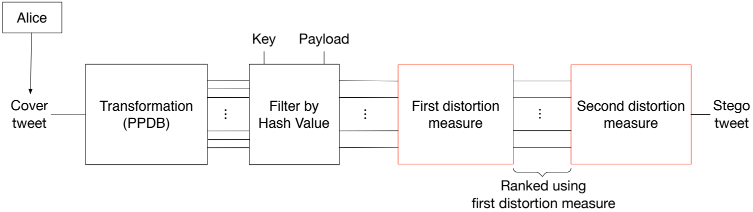
```
gosh today i truly don't want anything my beard to move away
        gosh now i genuinely don't want my beard to go away
   gosh now I genuinely just don't wanna my beard to go away
         gosh now I truly don't want my beard of going away
```



CoverTweet

```
gosh today i truly don't want anything my beard to move away
gosh now i genuinely don't want my beard to go away
gosh now I genuinely just don't wanna my beard to go away
gosh now I truly don't want my beard of going away
```

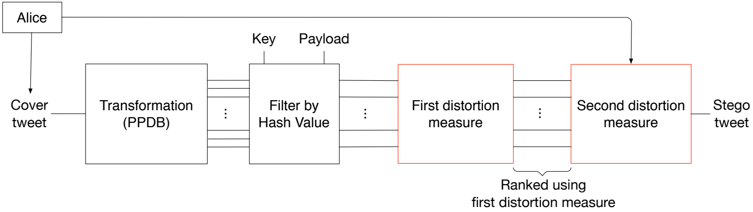
```
gosh now i genuinely don't want my beard to go away
gosh now I genuinely just don't wanna my beard to go away
gosh now I truly don't want my beard of going away
gosh today i truly don't want anything my beard to move away
```



CoverTweet

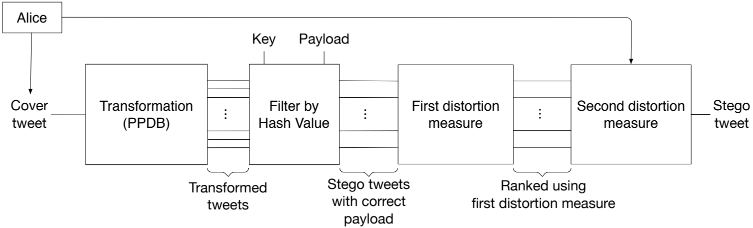
gosh today i truly don't want anything my beard to move away
gosh now i genuinely don't want my beard to go away
gosh now I genuinely just don't wanna my beard to go away
gosh now I truly don't want my beard of going away

gosh now i genuinely don't want my beard to go away
gosh now I genuinely just don't wanna my beard to go away
gosh now I truly don't want my beard of going away
gosh today i truly don't want anything my beard to move away



CoverTweet

gosh now I really don't want my beard to go away ⇨ gosh now i genuinely don't want my beard to go away



Statistical Machine Translation

- Model the probability that a stego sentence s is a translation of cover sentence c ($\Pr(s|c)$).
- Bayes' law:

$$\Pr(s|c) = \frac{\Pr(c|s)\Pr(s)}{\Pr(c)}$$

Statistical Machine Translation

- Model the probability that a stego sentence s is a translation of cover sentence c ($\Pr(s|c)$).
- Bayes' law:

$$\Pr(s|c) = \frac{\Pr(c|s)\Pr(s)}{\Pr(c)}$$

Statistical Machine Translation

- Model the probability that a stego sentence s is a translation of cover sentence c ($\Pr(s|c)$).
- Bayes' law:

$$\Pr(s|c) = \frac{\Pr(c|s)\Pr(s)}{\Pr(c)}$$

Language Modelling

- Our stego sentence s is made up of words $w_1, \ldots, w_T$.

$$
\begin{aligned}
\Pr(w_1, \ldots, w_T) &= \Pr(w_1) \prod_{i=2}^{T} \Pr(w_i | w_1, \ldots, w_{i-1}) \\
&\approx \Pr(w_1) \Pr(w_2 | w_1) \prod_{i=3}^{T} \Pr(w_i | w_{i-1}, w_{i-2})
\end{aligned}
$$

- This is a 2nd order Markov model

Language Modelling

- These probabilities are calculated using the maximum likelihood estimation (MLE):

$$\text{Pr}(\text{sat}|\text{the}, \text{cat}) = \frac{\text{count}(\text{the cat sat})}{\text{count}(\text{the cat})}$$

- Counts gathered from large text corpora (here 72M tweets). In practice, the counts are smoothed to avoid probabilities of 0.

Alice's Language Model

- What data can we use to train the language model?
- We need to train on cover data, of which we don't have enough of (a few hundred from Alice).
- We do have a huge amount of other twitter data (500M per day!).
- This is the problem of language model *adaptation*.

Alice's Language Model

- We train a small model on Alice's data, and a large model on general twitter data.
- The probabilities from both models are then linearly interpolated. For example:

$$\Pr(w_3|w_2, w_1) = (1-\lambda)\Pr_A(w_3|w_2, w_1) + \lambda \Pr_G(w_3|w_2, w_1)$$

Linguistic Distortion Measure

$$D(c, s) = -\log\left(\frac{\Pr(s|c)}{\Pr(c|c)}\right)$$

$$0 \leq D \leq \infty$$

Cover:

`I wish I was drinking a mojito right about now #keepingitreal`

Possible stego tweets:

1. `i wish i was drinking a mojito law around now #keepingitreal 0.815`
2. `i wish i was drink a mojito good about now #keepingitreal 1.229`
3. `if only i used to be drinking a mojito right about now #keepingitreal 1.670`
4. `i wish i was drinking a mojito right about far #keepingitreal 1.732`
5. `i 'd like to be drinking a mojito right around now #keepingitreal 1.878`

⋮

3000. `i wish i went drinkable a mojito entitled around today #keepingitreal 18.199`

Secondary Distortion Measure: Human Interaction

- Language modelling isn't good enough to guarantee that the option with lowest distortion is *actually* the best.
- Alice can choose the *true* best choice, from the ranked stego objects given by the first distortion measure.
- What if no option is fluent?
  - Alice can't signal no payload.
  - Recipient can't tell when there are no good options.
  - Alice will have to rewrite tweet, or not use it.

Evaluation

- Gathered 72M tweets, all posted in May 2013, from the Harvard TweetMap.
- Randomly selected 10 users with 'typical' characteristics:
  - Average number of words per tweet (11)
  - Average size of vocabulary per tweet (6)
  - Sufficient tweets for training the LM (> 500)
- For these 10 users, we trained the LM on the majority of their tweets.

Evaluation Data

Which of these have a hidden message?

1. `i just want to get above and beyond the state of drunk`
2. `I want someone to come on long midnight walks please, that be perfect`
3. `in fact i just need a pet tortoise in my life`
4. `sneaking your favourite foods into the trolley when you go food shopping`

Evaluation Data

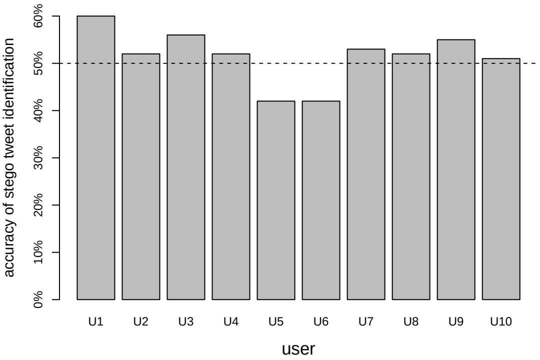
Which of these have a hidden message?

1. `i just want to get above and beyond the state of drunk`
2. `I want someone to come on long midnight walks please, that be perfect`
3. `in fact i just need a pet tortoise in my life`
4. `sneaking your favourite foods into the trolley when you go food shopping`

- 20 human judges were shown 80 innocent tweets from each user, followed by 20 unclassified tweets (10 stego, 10 innocent).
- The judges were asked: which of these are steganographic?

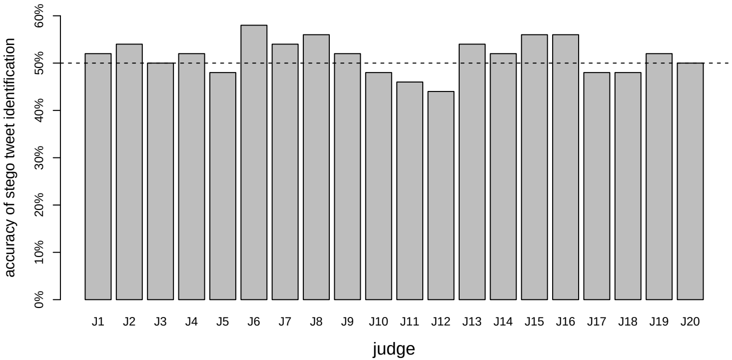
Results

- Of 1000 steganographic tweets shown to judges, only 515 were correctly identified.
- Insufficient evidence to reject the null hypothesis that judges are guessing randomly



Results

- Of 1000 steganographic tweets shown to judges, only 515 were correctly identified.
- Insufficient evidence to reject the null hypothesis that judges are guessing randomly.



s

Summary

- Very secure against human judges!
- Embedded more bits per stego object than existing systems.
- More work needed on statistical detection methods.
- Synchronisation and coding need considerable work.