ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Near-Imperceptible Neural Linguistic Steganography via Self-Adjusting Arithmetic Coding

Jiaming Shen, Heng Ji, Jiawei Han

University of Illinois at Urbana-Champaign

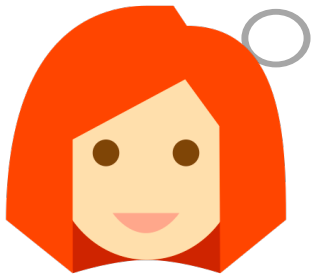
Presented by Jiaming Shen @ EMNLP 2020

Paper Link: https://arxiv.org/abs/2010.00677
Code & Data: https://github.com/mickeystroller/StegaText

Linguistic Steganography

Secret Plaintext

2am at Room 9112B, pwd abcxyz



Alice
(Sender)

Encode

Public Channel

Cover text

Bipartisan bill would requires a $13 billion appropriations at the end of the current fiscal year, Under the Act…

Eavesdropper

Bob
(Receiver)

Decode

“2am at Room 9112B, pwd abcxyz” get it!!!

Past Work

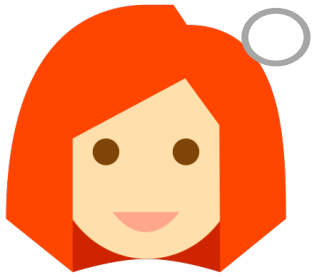
- **Edit-based methods** try to edit the secret plaintext and transform it into the innocent cover text
  - Synonym substitution (Topkara et al. 2006), Paraphrase substitution (Chang and Clark, 2010), Syntactic transformation (Safaka et al. 2016)
  - Cannot encode information concisely
- **Generation-based methods** aim to directly output the cover text by generating a series of words based on a language model (LM)
  - Bin-LM (Fang et al. 2017), RNN-Stega (Yang et al. 2019), Patient-Huffman (Dai and Cai, 2019), Arithmetic (Ziegler et al. 2019)
  - Better coding algorithms with modern LMs lead to better performance

Our method fails in the “generation-based” category

Formalization: Steganography Pipeline

Secret Plaintext

$\mathbf{m} \in \{0,1\}^L$



Alice
(Sender)

Encode
$f(\cdot)$

Public Channel

Cover text

$\mathbf{y} = f(\mathbf{m})$

Eavesdropper

Bob
(Receiver)

Decode

$\mathbf{m} = f^{-1}(\mathbf{y})$

Formalization: Imperceptibility

- Following (Dai and Cai, 2019), we define the "imperceptibility" of a steganography algorithm based on the Total Variation Distance (TVD):
  - The true language distribution $\mathbf{P}^*_{LM}$
  - The cover text distribution $\mathbf{Q}(\mathbf{y})$ implicitly defined by the encoder $f(\cdot)$

$$\mathrm{TVD}(\mathbf{P}^*_{LM}, \mathbf{Q}) = \frac{1}{2}\|\mathbf{Q} - \mathbf{P}^*_{LM}\|_1$$

$$\leq \frac{1}{2}\|\mathbf{Q} - \mathbf{P}_{LM}\|_1 + \frac{1}{2}\|\mathbf{P}_{LM} - \mathbf{P}^*_{LM}\|_1$$

Measures how good the encoder is

Measures how good the LM is

Toward “Near-Imperceptible”

- Decompose sequence-level imperceptibility to per-step imperceptibility:

$$\frac{1}{2}\|\mathbf{Q}-\mathbf{P}_{LM}\|_1 \leq \sqrt{\frac{\ln 2}{2} D_{KL}(\mathbf{Q}\|\mathbf{P}_{LM})}$$

$$= \sqrt{\frac{\ln 2}{2}\sum_{t=1}^{\infty} D_{KL}(\mathbf{Q}(\cdot|\mathbf{y}_{<t})\|\mathbf{P}_{LM}(\cdot|\mathbf{y}_{<t}))}$$

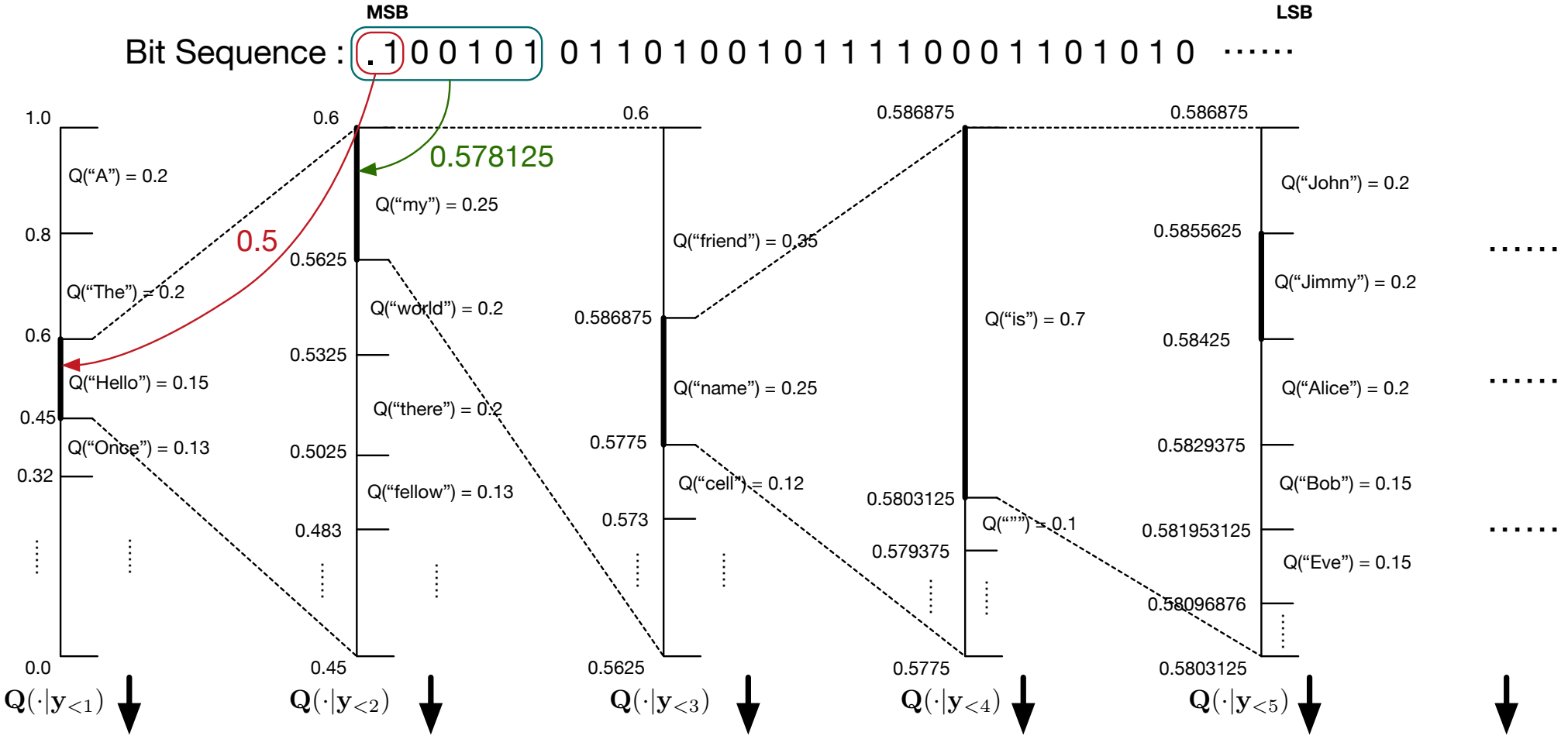
KL divergence of each generation step

If each step’s KL divergence is small enough, the overall generated cover text is statistically “near-imperceptible”

Background: Arithmetic Coding

- View the bit sequence **m** as a binary fractional number
  - For example: $\mathbf{m} = [1, 0, 1] \rightarrow B(\mathbf{m}) = 0.101 \rightarrow 1\times2^{-1} + 1\times2^{-3} = 0.625$
- At the initial time step $t = 1$, create an interval $[l_1, u_1) = [0, 1)$
- Each time step t starts with $[l_t, u_t)$ and outputs a token y_t by:
  - Compute the conditional distribution $\mathbf{Q}(y_t|\mathbf{y}_{<t})$
  - Divide the current interval $[l_t, u_t)$ into sub-intervals, each representing a fraction of the current interval proportional to $\mathbf{Q}(y_t|\mathbf{y}_{<t})$ of a possible next token
  - Select the sub-interval that $B(\mathbf{m})$ lays in and set it to be $[l_{t+1}, u_{t+1})$
  - Output the token corresponding to the above selected sub-interval
- Stop when all **m**-prefixed fractions fall into the final interval

Arithmetic Coding Example



How to Get Cover Text Distribution Q?

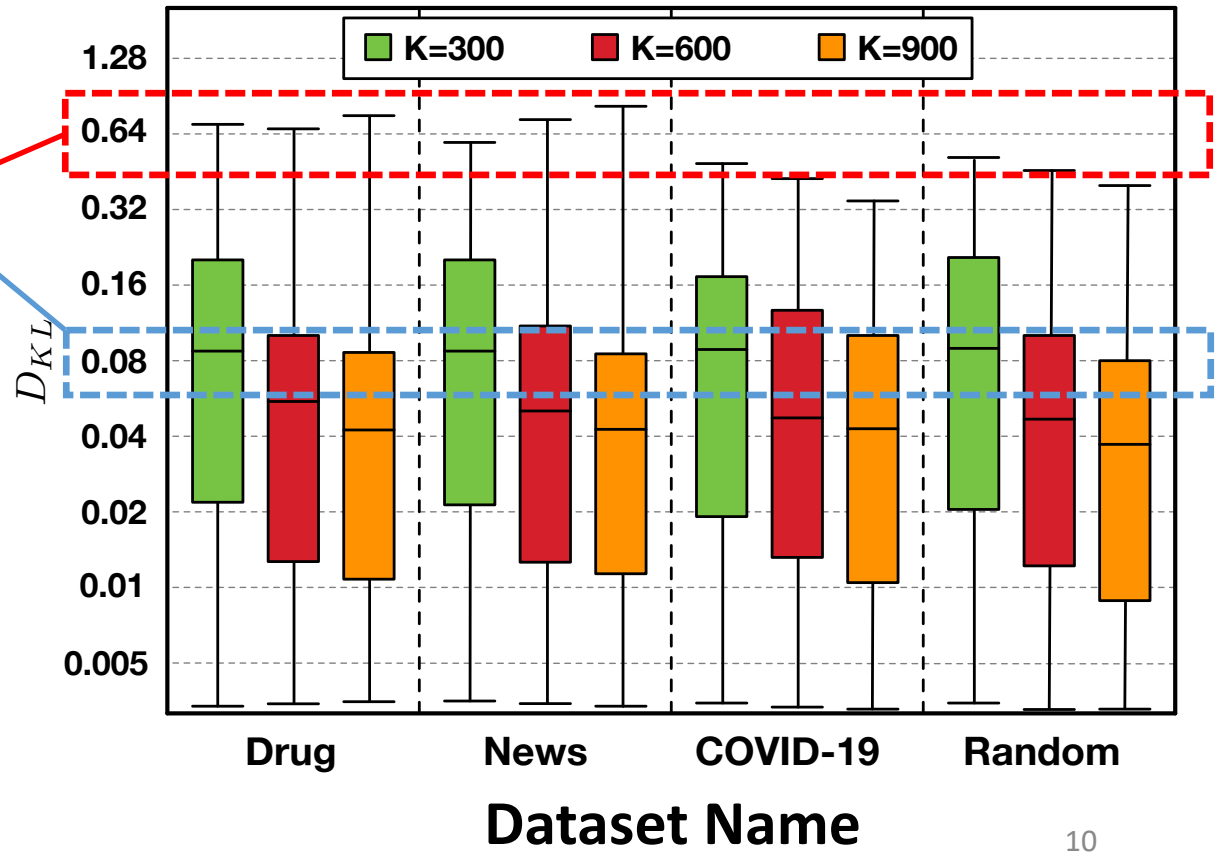
- Directly use $\mathbf{P}_{LM}$?
  - Generate tokens with low probabilities -> bad cover text quality
  - Lead to the precision issue and very slow encoding/decoding speed
- (Static) *top K* sampling (Ziegler et al. 2019):
  - Select *K* most likely tokens based on $\mathbf{P}_{LM}$
  - *K* is predefined and **fixed cross all steps**

$$\mathbf{Q}(\mathbf{y}_t|\mathbf{y}_{<t}) \propto \begin{cases} \mathbf{P}_{LM}(\mathbf{y}_t|\mathbf{y}_{<t}) & \text{if } \mathbf{y}_t \in \text{argtopK}_{\mathbf{y}'}\ \mathbf{P}_{LM}(\mathbf{y}'|\mathbf{y}_{<t}) \\ 0 & \text{otherwise} \end{cases}$$

Limitation of Static Top *K* Sampling Strategy

- Large variations across different steps

While the median imperceptibility is reasonable, the maximum imperceptibility is intolerable



Per-step imperceptibility:

$$D_{KL}(\mathbf{Q}(\cdot|\mathbf{y}_{<t})\|\mathbf{P}_{LM}(\cdot|\mathbf{y}_{<t})) = -\log Z$$

$$Z = \sum_{\mathbf{y}' \in \text{argtop}K \ \mathbf{P}_{LM}(\mathbf{y}'|\mathbf{y}_{<t})} \mathbf{P}_{LM}(\mathbf{y}'|\mathbf{y}_{<t})$$

Self-Adjusting Arithmetic Coding

- Key idea: in each step, choose the smallest *K* that achieves the imperceptibility requirement

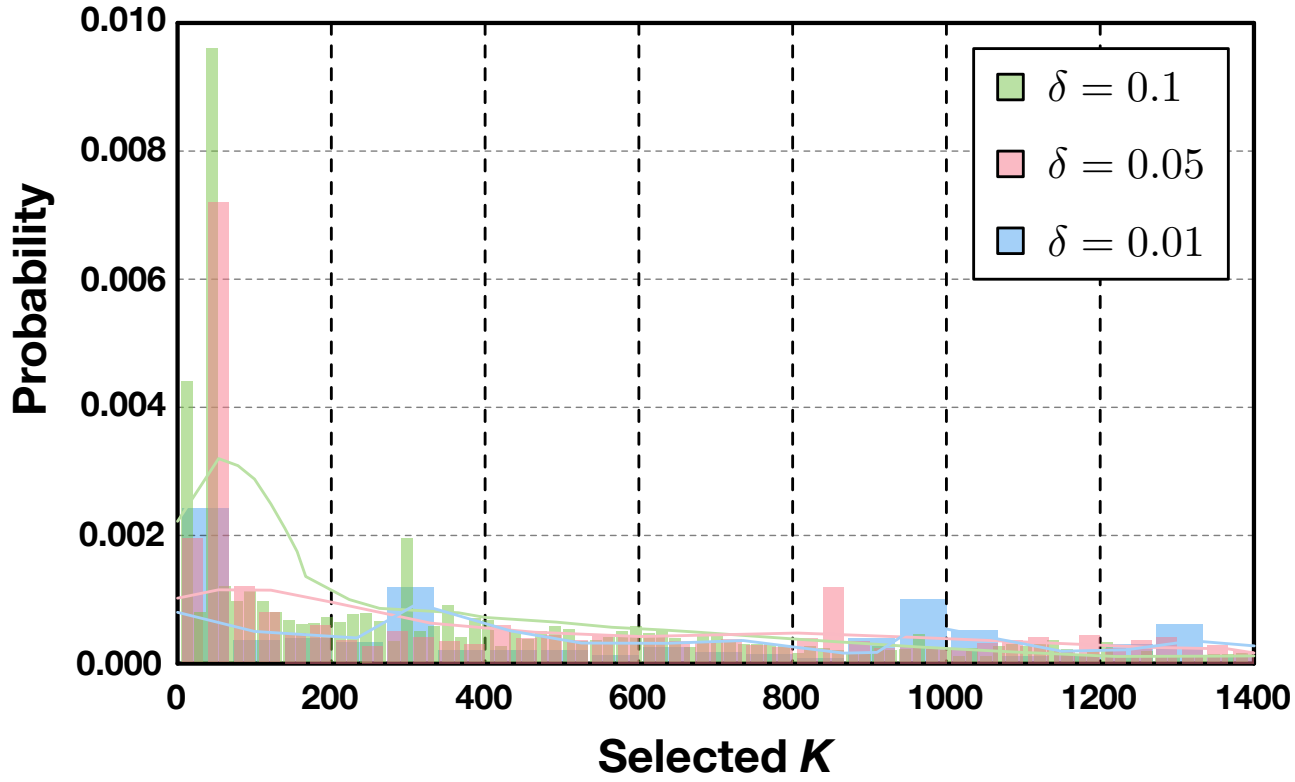
$$D_{KL}(\mathbf{Q}(\cdot|\mathbf{y}_{<t})\|\mathbf{P}_{LM}(\cdot|\mathbf{y}_{<t})) \leq \delta$$

Imperceptibility requirement

$$K^* = \min(\{K| \sum_{\mathbf{y}' \in \text{argtop}K \ \mathbf{P}_{LM}(\mathbf{y}'|\mathbf{y}_{<t})} \mathbf{P}_{LM}(\mathbf{y}'|\mathbf{y}_{<t}) \geq e^{-\delta}\})$$

Selected *K* Distribution

- In majority of steps, choosing $K < 50$ is enough
- Many steps do require large $K > 1000$



Experiments

Dataset	Drug	News	COVID-19	Random
Num. of Sentences	3972	6437	2000	3000
Avg. Num. of Words	19.01	14.30	24.21	—
Avg. Num. of Bits	289.75	211.08	308.65	256

- Datasets:
  - **Drug**: Reddit drug comments
  - **News**: the 4th sentences in CNN/DM news articles
  - **COVID-19**: the 4th sentences in paper abstracts
  - **Random**: 256-bit sequences generated following the uniform distribution
- Metrics:
  - **Bits/word**: Average number of message bits encoded by a cover text token
    - Larger is better (meaning the information is encoded concisely)
  - **KL**: The KL divergence between cover text distribution and LM distribution
    - Smaller is better (meaning the generated cover texts are natural and faithful)

Overall Results

- Arithmetic coding methods are better than Huffman tree based methods

Methods	Drug		News		COVID-19		Random	
	Bits/Word ↑	D_{KL} ↓	*Bits/Word* ↑	D_{KL} ↓	*Bits/Word* ↑	D_{KL} ↓	*Bits/Word* ↑	D_{KL} ↓
Bin-LM ($B = 1$)	1	1.864	1	1.922	1	1.838	1	1.185
Bin-LM ($B = 2$)	2	2.358	2	2.385	2	2.346	2	2.374
Bin-LM ($B = 3$)	3	2.660	3	2.680	3	2.659	3	2.664
RNN-Stega ($H = 3$)	2.370	1.015	2.387	1.015	2.368	0.999	2.378	0.991
RNN-Stega ($H = 5$)	3.399	0.628	3.393	0.628	3.368	0.624	3.370	0.630
RNN-Stega ($H = 7$)	4.202	0.424	4.202	0.426	4.197	0.426	4.163	0.422
Patient-Huffman ($\epsilon = 0.8$)	1.835	0.269	1.834	0.269	1.844	0.270	1.847	0.271
Patient-Huffman ($\epsilon = 1.0$)	2.147	0.360	2.154	0.361	2.142	0.357	2.148	0.358
Patient-Huffman ($\epsilon = 1.5$)	2.596	0.524	2.583	0.522	2.579	0.519	2.584	0.520
Arithmetic ($K = 300$)	3.497	0.203	3.491	0.209	3.510	0.191	3.466	0.189
Arithmetic ($K = 600$)	4.247	0.162	4.240	0.166	4.289	0.146	3.599	0.160
Arithmetic ($K = 900$)	4.376	0.149	4.358	0.152	4.414	0.131	3.669	0.147
SAAC ($\delta = 0.1$)	4.262	0.153	4.232	0.157	4.301	0.133	4.225	0.136
SAAC ($\delta = 0.05$)	4.451	0.134	4.441	0.138	4.519	0.114	4.419	0.117
SAAC ($\delta = 0.01$)	**4.862**	**0.109**	**4.784**	**0.117**	**4.851**	**0.093**	**4.778**	**0.099**

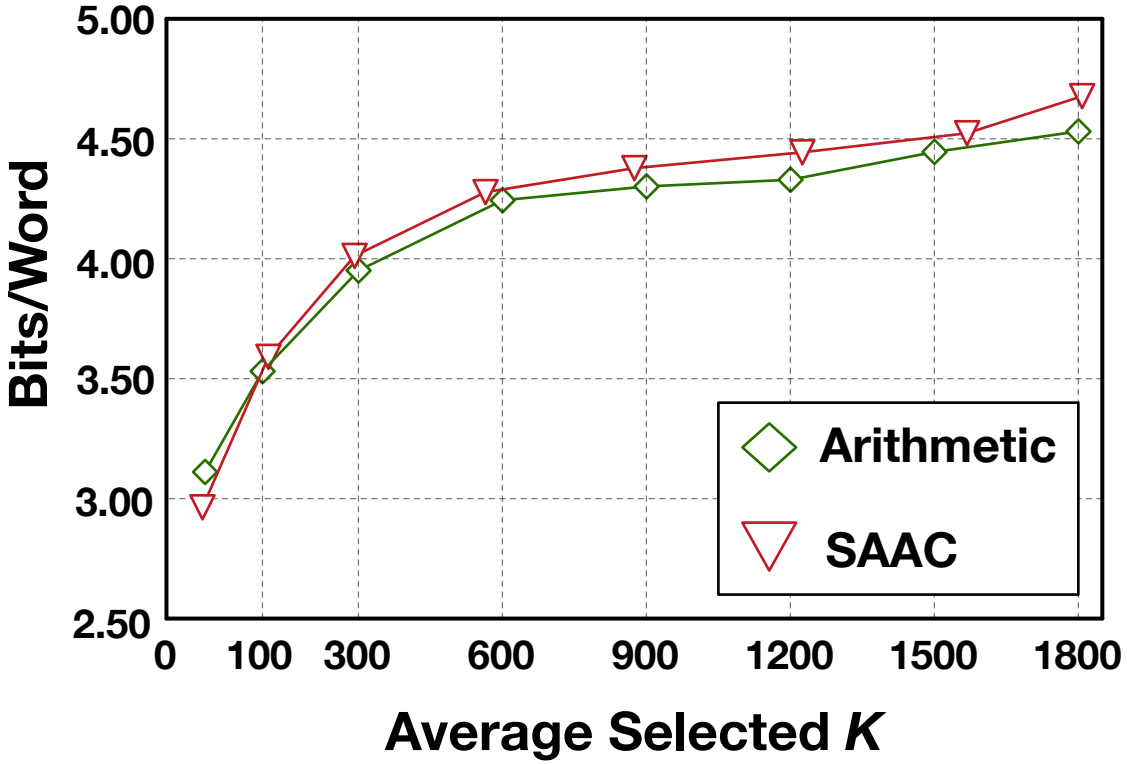
Overall Results

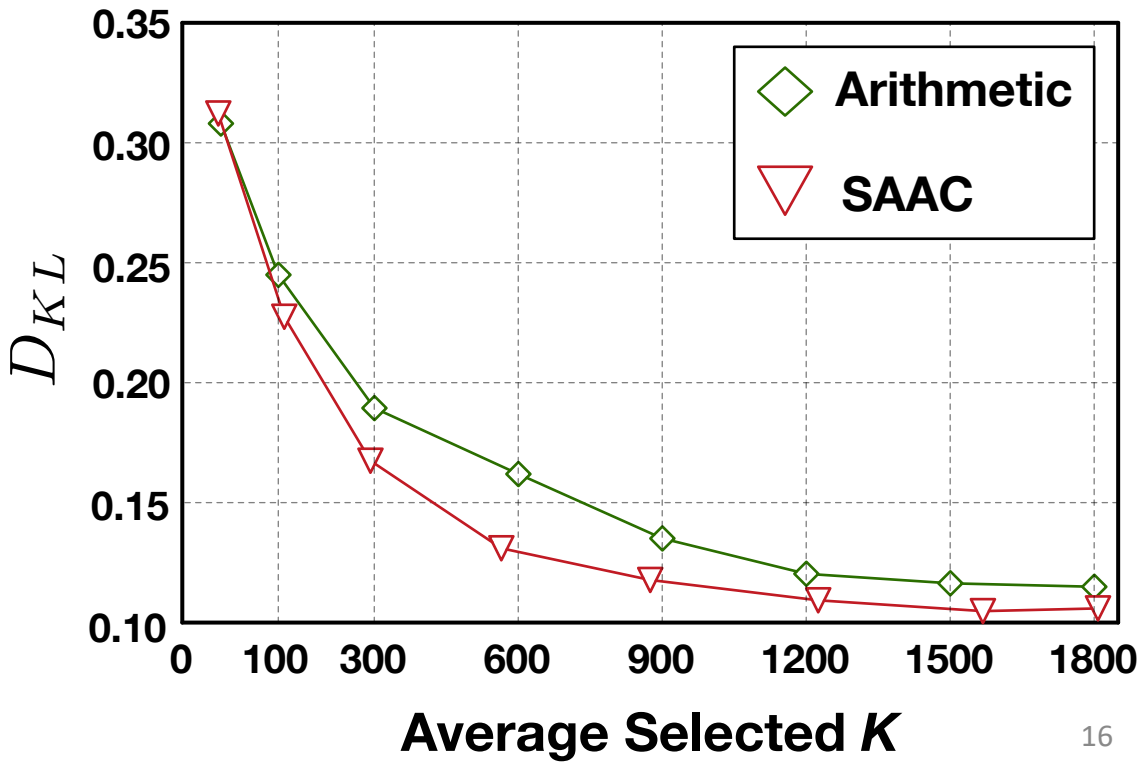
- Our proposed Self-Adjusting Arithmetic Coding (SAAC) method is better than the static Arithmetic Coding method

Methods	Drug		News		COVID-19		Random	
	Bits/Word ↑	D_{KL} ↓	*Bits/Word* ↑	D_{KL} ↓	*Bits/Word* ↑	D_{KL} ↓	*Bits/Word* ↑	D_{KL} ↓
Bin-LM ($B = 1$)	1	1.864	1	1.922	1	1.838	1	1.185
Bin-LM ($B = 2$)	2	2.358	2	2.385	2	2.346	2	2.374
Bin-LM ($B = 3$)	3	2.660	3	2.680	3	2.659	3	2.664
RNN-Stega ($H = 3$)	2.370	1.015	2.387	1.015	2.368	0.999	2.378	0.991
RNN-Stega ($H = 5$)	3.399	0.628	3.393	0.628	3.368	0.624	3.370	0.630
RNN-Stega ($H = 7$)	4.202	0.424	4.202	0.426	4.197	0.426	4.163	0.422
Patient-Huffman ($\epsilon = 0.8$)	1.835	0.269	1.834	0.269	1.844	0.270	1.847	0.271
Patient-Huffman ($\epsilon = 1.0$)	2.147	0.360	2.154	0.361	2.142	0.357	2.148	0.358
Patient-Huffman ($\epsilon = 1.5$)	2.596	0.524	2.583	0.522	2.579	0.519	2.584	0.520
Arithmetic ($K = 300$)	3.497	0.203	3.491	0.209	3.510	0.191	3.466	0.189
Arithmetic ($K = 600$)	4.247	0.162	4.240	0.166	4.289	0.146	3.599	0.160
Arithmetic ($K = 900$)	4.376	0.149	4.358	0.152	4.414	0.131	3.669	0.147
SAAC ($\delta = 0.1$)	4.262	0.153	4.232	0.157	4.301	0.133	4.225	0.136
SAAC ($\delta = 0.05$)	4.451	0.134	4.441	0.138	4.519	0.114	4.419	0.117
SAAC ($\delta = 0.01$)	**4.862**	**0.109**	**4.784**	**0.117**	**4.851**	**0.093**	**4.778**	**0.099**

Detailed Comparisons with Arithmetic Baseline

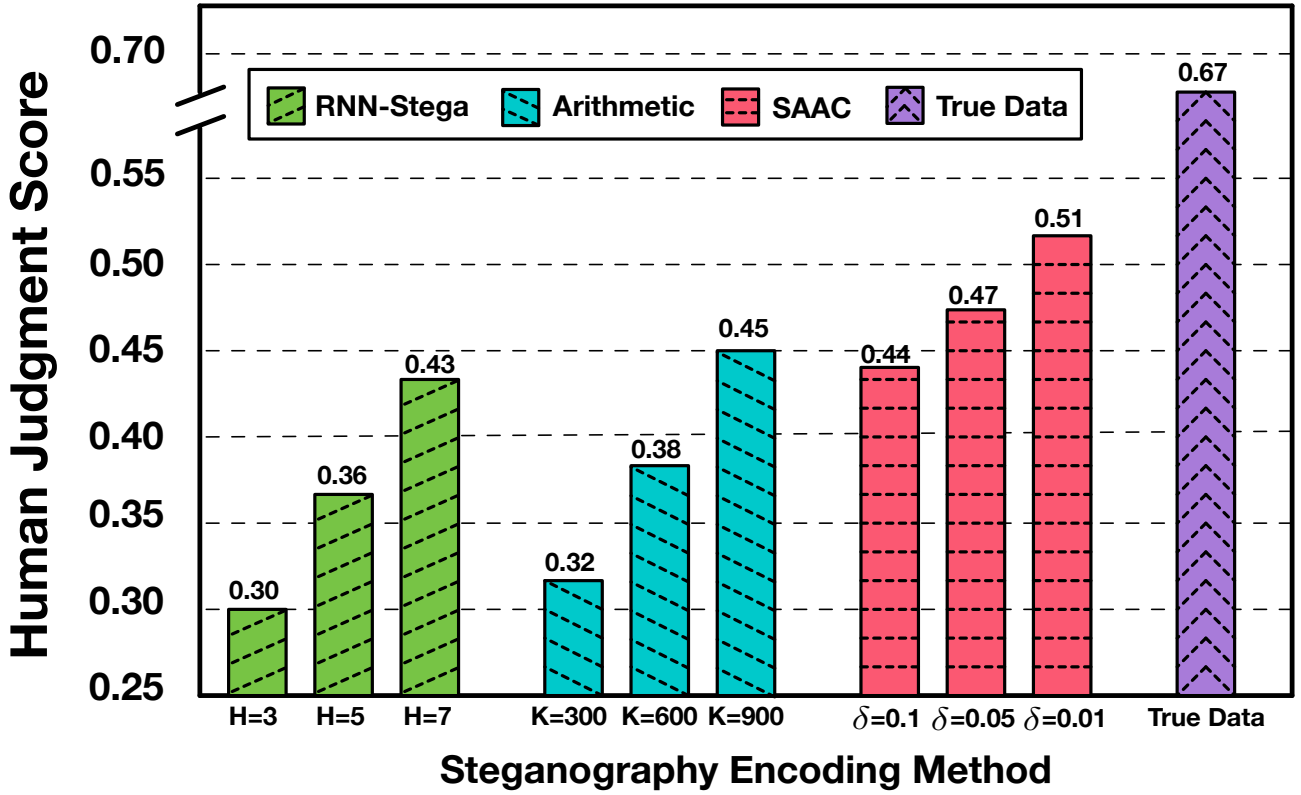
- Our new algorithm achieves the better imperceptibility (smaller D_{KL}) and encodes secret messages more concisely (larger Bits/Word)





Human Evaluation

- Can fool human eavesdroppers more than half of the times



Case Study: Generated Cover Text

- Generated cover texts are fluent and can effectively hide the information in the original plaintexts

Plaintext	Context	Generated Cover Text
Phylogenetic analysis showed that Bat-SARS-CoV formed a distinct cluster with SARS-CoV. <eos>	The first tally is in. The HEROES Act, passed the House of Representatives Friday evening, would reduce federal revenue by a net total of \$883 billion between 2020 and 2030, according to the Joint Committee on Taxation (JCT). It is highly unlikely that the bill will get signed into law as is, given the White House's veto threat and Senate Republican's view of it as hardly salvageable.	bipartisan bill would require a \$13 billion appropriation at the end of the current fiscal year. Under the Act, you would not collect federal taxes on drugs or make drug-related appropriation if you were a major manufacturer of cannabis. The proposal will likely give Trump the opportunity to only fund the 10 types of confiscated marijuana that the federal government has been conducting a current drug .
molly ultra caps capped at 180mgs will have you flying for hrs clean come down 99 of the time . <eos>	Washington received his initial military training and command with the Virginia Regiment during the French and Indian War. He was later elected to the Virginia House of Burgesses and was named a delegate to the Continental Congress, where he was appointed Commanding General of the nation's Continental Army. Washington led American forces, allied with France, in the defeat of the British at Yorktown.	Confederate troops were assigned a plaque near Berrien\'s Mill, a creek south of New York City. His monument of Lafayette\'s power to bear arms became the very flag of the Union government. Washington returned to Pennsylvania in 1788 when the navy introduced Continental forces to Britain. Five years later the \"Black Ships\" were commissioned

Conclusion and Future Work

- In this study, we
  - Show arithmetic coding based methods are near imperceptible
  - Propose a new steganography method that achieves “near-imperceptibility”
- In the future, we plan to
  - Analyze more factors in the steganography algorithm, including
    - Different LMs (GPT-2 small/medium/large/x-large, XLNet?)
    - Different ways to convert natural language plaintext to bit-sequence plaintext
  - Use learning techniques to find the best *K*
  - Improve the encoding/decoding speed

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Thanks for your attention

Questions ❓

Email: js2@illinois.edu

Paper Link: https://arxiv.org/abs/2010.00677
Code & Data: https://github.com/mickeystroller/StegaText