1. **"All Your Secret Data Makes a Great Nigerian Prince"**
   Encoding data for quantitative and qualitative properties through greedy algorithms for no reason at all
2. **Some words/phrases for context**
   - Markov text generator
     - Dissociated press
     - Mark V Shaney
       - *"I spent an interesting evening recently with a grain of salt."*
   - Probability distribution function
     - Cumulative distribution function
     - Quantile function
     -
   - Zipf's Law

     -
3. **Motivation**
   - DNS tunnel detection
     - Assumption: DNS queries have a characteristic character frequency distribution
     - Assumption: DNS tunnels produce a character frequency distribution that is quantifiably different from that of normal queries.
   - Several detection methods rely on these two assumptions.
4. **It is true!**
5. **It is true!**
6. **It is also true!**
7. **It is also true!**
8. **... But what if it weren't?**
   - Then these detection methods would fail miserably.
   - Is it possible to manipulate arbitrary data into a form that possesses certain statistical properties?
9. **Relation to Statistics**
   -
   - Can be thought of as a numerical quantile function of sorts
     - Quantile functions take in uniform random data and output random data that conforms to a distribution.
   -
10. **Theoretically**
    - Yes
    -
    - Specify the PDF description of the desired output, and then Shannon can come into play.
      - If the output entropy is lower, we will see inflation of the stream
      -
11. **In Practice**
    - As it turns out, also yes.
    -
    - 1422 lines of C and 123 lines of Bash implements character distribution matching as well as Markovian text generation and some output prettifying.
12. **So does it really work?**
13. **So does it really work?**

14 **So does it really work?**

15 **So does it really work?**

teiocrgbahnuswmk doejtael ev npfsiiaeyqxtsaldhewoureis nrtazlreaeeiishcru po atdelisemeyia toanrlershe uvhgoigfw onatscr bentadlem topieensa dhrft istna luderephnwyts a ieiemtaongclr edtaoroueibecwka sl ieenhvoigueyt ranosdta plseirore tihe etfuda gwie renmhnsiethf sctnraeyroslwaiosdupa lhaehemriaocge stnhe wtsasb avfeloaelid

16 **Markovian Text Generation**

- Since a Markov chain can be thought of as a PDF where each 'symbol' is a PDF, Markov chains were easily implemented.
-
- Generated the transition data from Sherlock Holmes, Huckleberry Finn, the complete works of Alexander Dumas, and Dante's Divine Comedy.

17 **Feed it random data…**

- 25 bytes, one-word transition matrix.
-

Other, that the king he was a little, but I was the way, and the time. The world, which I have a man. But the rest of a good. I could see the matter, when I am sure that, as the last, I see it was, the door, who had a long as I had been a couple of them. You have to be a lot of the first. In the right, where I got to do not, so that you know what you think that I know that he, it would be the old gentleman, for a few minutes, this is.

18 **Feed it random data…**

- 25 bytes, two-word transition matrix.
-

The king, who was the first time, and the two men, who has been a very good. And what did you not? Yes. Well? Asked dartagnan, and I am a man of the young girl. I am the only one of the church, a little. I have been the case of the two friends. I will not be able, he had not been able to see the king. I was not a man in a low tone, and a half, also, said he, in a few minutes the belfry, where the devil! Cried the king was at this time the door, and who was in the morning. The young men, but the queen, and that the queen of england. He had been made to the queen. I had been taken from me, and he was a great number of the world, I have not been a soldier. And now.

19 **Feed it random data…**

- 10 bytes, three-word transition matrix.
-

And the other two dying. The first was the wish to make me a little, and colored a great deal, it seems to me that you are a brave man, and I will tell him that I shall never be aught but an ass, answered la ramee. Grimaud, in a voice of gloom what have I done to you, and I have no more to be dreaded, and displayed his own friendship for the victim, as the king has said nothing to porthos of his wound, as usual, and the two friends were announced they started and exchanged some rapid.

20 **How does it fare?**

21 **How does it fare?**

22 **How does it fare?**

23 **Summary**

- The initial motivation:
  - Evade detection based on character frequency

–
- The outcome:
  - Able to encode/decode arbitrary data into a form that is nearly indistinguishable from English to a computer

24 **Why?**
-
-
-
- What practical use does this have?
-

25 **Possibility:**
Covert Instruction Spam
- How could you send secret data to an operative in the field, without anyone knowing you sent anything to them?
  - Encode the covert data as spammy English, and send it to millions of people.
  - How would you identify the intended recipient?

26 **Intermission**
- High level overview complete
- Normal people can leave
-
- What follows:
  - Implementation discussion
  - Security discussion
  - Technical details

27 **Implementation Overview:**
Inputs
- Input: High-entropy source
  - Easily obtained via compression or encryption
- Input: Description of PDF to match

28 **Implementation Overview:**
Setup
- Sort PDF from most to least probable symbol
- Map between symbols and bit strings
  - Prefix-free mappings are not a requirement here (you'll see why later)
  - Each symbol gets a bit string that is at least as likely in a high-entropy input as the symbol should be in the output
    - This ensures that we get enough opportunities to choose this as the output symbol

29 **Implementation Overview:**
Setup
- The two most common symbols get mapped to single bits 0 and 1
  - This ensures we can eat up extra bits with something
- Two ways of mapping bit strings:
  - Efficient: Choose the longest bit string that satisfies the opportunity requirement
  - Accurate: Choose the shortest bit string we haven't used yet

30 **Implementation Overview:**
Encoding (Greedy Algorithm)
- Look at the first N bits of input, where N is the maximum length of a bit string we mapped to a symbol

- Find the symbols that could match (a portion of) the bits
  - Choose the one that is farthest from reaching its quota so far (greedy portion). Resolve ties arbitrarily.

31 **Implementation Overview:**
Decoding
- Look at a symbol of input
- Output its associated bit string
  - Because parsing the input distribution is completely deterministic, so are bit string mappings

32 **Security**
- Unicity distance is weird here (Especially for Markov chains)
  - 'Key' may be way larger than message
- No consistent notion of keyspace size
- Large amounts of output by definition reveal the input distribution allowing for decoding
  - But do they reveal enough to be practically dangerous?
-
-

33 **Security:**
Convergence to Input Distribution

34 **Security:**
Convergence to Input Distribution
- Unicity distance of 6!
- Original
  - teiouhgbarscnmwk loejtaed ev sa
-
- 10-million character sample
  - teiocrgbahnuswmk doejtael ev na

35 **Security:**
Convergence to Input Distribution
- Systemic bias is introduced
- This is implementation specific behaviour though
  - Can this be adjusted for? Note the 100k-v-1M shift gives direction of "off"-ness.
  - Adjust everything we got from the 10-M sample by 3.5%.

36 **Security:**
Convergence to Input Distribution

37 **Security:**
Convergence to Input Distribution

38 **Security:**
Convergence to Input Distribution
- What is the unicity distance this time?
-
-
  -

39 **Security:**
Convergence to Input Distribution
- What is the unicity distance this time?
- 9899 characters of output!
-

- Original:
  - …at euieno amtwfpvldoshs iretcra enogorsdi watpyour
  -
- Adjusted
  - …at euieno amtwfpvldoshs iretcra enogorsdi watpybl

40 **Security:**
Convergence to Input Distribution

-
-
- Markov chains will have considerably more convoluted convergence properties
  -

41 **Security**

-
- This is not suitable as a replacement for a proper encryption!
-
- It may be suitable for steganographic applications though.

42