Vigenère's cipher, section 2.3

- Code the alphabet using 0 - 25: A - 0, B - 1, C - 2, ...
- We work with arithmetic modulo 26.
- This cipher encrypts strings of letters - we skip blanks. E.g.

  D O G
  3 14 6

- The cipher uses a code length k and a vector of length k of numbers mod 26.
- For example, if $k = 3$ and $v = (4, 7, 12)$ we encrypt DOG as follows:

$$(3, 14, 6) + (4, 7, 12) = (7, 21, 18)$$

  and that is the string HWT.

Vigenère's cipher, cont'd

- For longer strings we just code the first k letters as above and then start again with the next k letters until we finish the string.
- The sense of security comes from not knowing k as well as not knowing v.
- As we will see, this cipher is susceptible to a letter frequency attack.

English letter frequency

Letter frequency in texts, Beker-Piper, '82

a	b	c	d	e	f	g	h	i
.082	.015	.028	.043	.127	.022	.020	.061	.070
j	k	l	m	n	o	p	q	r
.002	.008	.040	.024	.067	.075	.019	.001	.060
s	t	u	v	w	x	y	z	
.063	.091	.028	.010	.023	.001	.020	.001	

Cracking Vigenère's cipher

- How do we break this cipher? The issue is finding the key length.
- You need to have a reasonably long chunk of ciphertext (long relative to the potential key length).
- Compare the ciphertext to a copy of the ciphertext displaced by ℓ places and count the number of spots with the same character.
- The number ℓ with the greatest number of matches is likely to be the key length. Why?

Proof that this works

- Suppose $V = (p_0, p_1, \ldots, p_{25})$ is a vector of the letter frequencies from the earlier slide.
- Consider the mth spot in the ciphertext which has been shifted by i by the cipher and then the displaced ciphertext which has been shifted by j. What are the chances that these two spots are the same character?
- $$p_{0-i}p_{0-j} + p_{1-i}p_{1-j} + \ldots p_{25-i}p_{25-j}$$

  where all the arithmetic is modulo 26.
- This is the same as

  $$p_0 p_{i-j} + p_1 p_{1+i-j} + \ldots p_{25} p_{25+i-j}.$$

Proof that this works, cont'd

- If we let V_i be the vector V shifted by i then the probability we just calculated is $V \cdot V_{i-j}$.
- By the Cauchy-Schwartz inequality, this is maximized when $i = j$.
- In fact $V \cdot V \approx .066$ and $V \cdot V_i \leq .045$ for $i \neq 0$.
- Once you have the key length k, you can consider the distribution of the kth letters to determine the actual key.
- Since this is a shift cipher, it suffices to just figure out what e is which should be the most frequent letter (appearing about 12.7% of the time).