JOHNS HOPKINS
UNIVERSITY

A Natural Language Approach to Automated Cryptanalysis of Two-time Pads

Joshua Mason
Kathryn Watkins
Jason Eisner
Adam Stubblefield

The Two Time Pad Problem

Attack at Dawn $\oplus$ doQvYcSWIPyXaC

Attack at Dawn ⊕ doQvYcSWIPyXaC

Take the Beach ⊕ doQvYcSWIPyXaC

Attack at Dawn ⊕ doQvYcSWIPyXaC

⊕ Take the Beach ⊕ doQvYcSWIPyXaC

Attack at Dawn ⊕ doQvYcSWIPyXaC

⊕ Take the Beach ⊕ doQvYcSWIPyXaC

Attack at Dawn doQvYcSWIPyXaC

⊕ ⊕ ⊕

Take the Beach doQvYcSWIPyXaC

Attack at Dawn ⊕ doQvYcSWIPyXaC

⊕ Take the Beach ⊕ doQvYcSWIPyXaC

Attack at Dawn ~~doQvYcSWIPyXaC~~

⊕ ⊕ ~~⊕~~

Take the Beach ~~doQvYcSWIPyXaC~~

Attack at Dawn ⊕ doQvYcSWIPyXaC

⊕ Take the Beach ⊕ doQvYcSWIPyXaC

Attack at Dawn

⊕

Take the Beach

`Attack at Dawn`

$\oplus$ = 15 15 1f 04 43 1f 48 04 54 62 21 00 14 6

`Take the Beach`

LINKSYS

LINKSYS

LINKSYS

OJNcDfoMncXzYwwQQZRXYWORT190LP

OJNcDfoMncXzYwwQQZRXYWORT190LP

⊕ the

```
  OJNcDfoMncXzYwwQQZRXYWORT190LP
⊕ the
----------------------------------
  QpL
```

```
OJNcDfoMncXzYwwQQZRXYWORT190LP
```

⊕ **the**

```
    OJNcDfoMncXzYwwQQZRXYWORT190LP
⊕    the
----------------------------------
     Man
```

Formalized by F. Rubin in 1978

Automated by E. Dawson and L. Nielson in 1996

Assumptions

- Uppercase English characters and space
- Space is always the most frequent character

$P_0 \oplus P_1$ = 6e 71 00 6f 79 61

$P_0 \oplus P_1 =$ 6e 71 00 6f 79 61

$P_0 \oplus P_1 = $ `6e 71      6f 79 61`

$P_0 \oplus P_1 =$ 6e 71 6f 79 61

$P_1 \oplus P_2 = \texttt{67 82 00 00 00 00 00 34}$

$P_1 \oplus P_2 =$ 67 82 00 00 00 00 00 34

$P_1 \oplus P_2 =$ 67 82 00 00 00 34

Testing Methodology

- Trained on the first 600K characters of the Bible
- Attempted recovery of passages from first 600K characters of the bible

Percentage Correctly Recovered

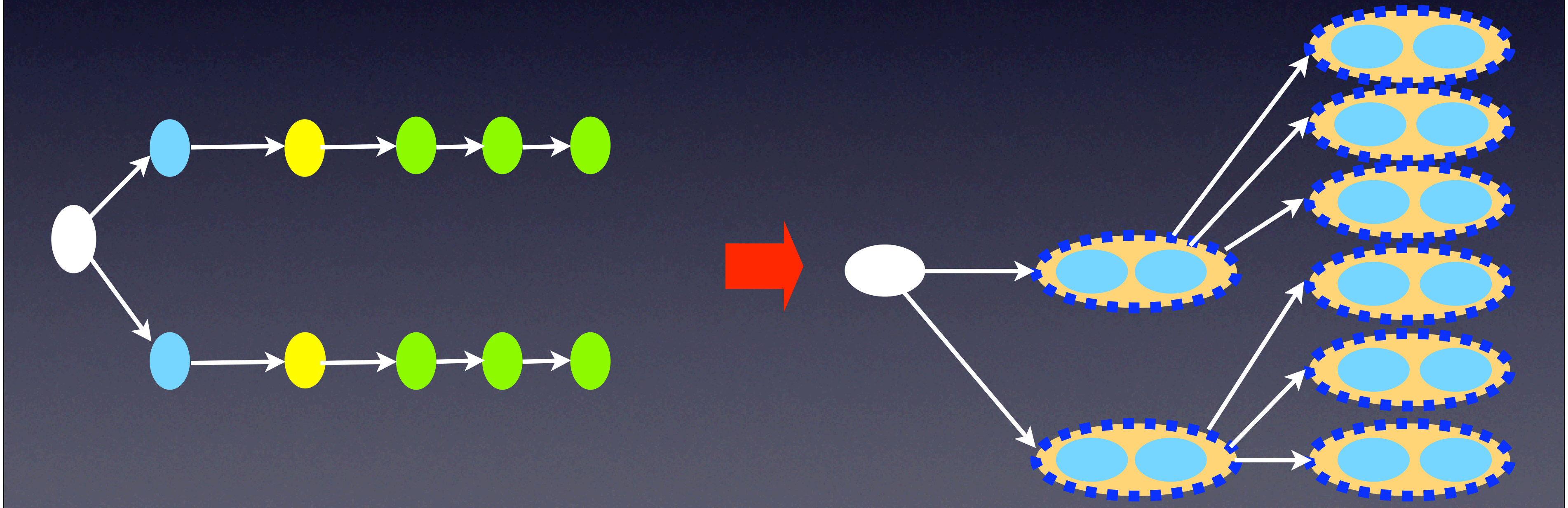
	Dawson & Nielson	
$P0 \oplus P1$	62.7%	
$P1 \oplus P2$	61.5%	
$P0 \oplus P1$	62.6%	

	Percentage Correctly Recovered	
	Dawson & Nielson	Our Technique
P0 ⊕ P1	62.7%	100%
P1 ⊕ P2	61.5%	99.99%
P0 ⊕ P1	62.6%	99.96%

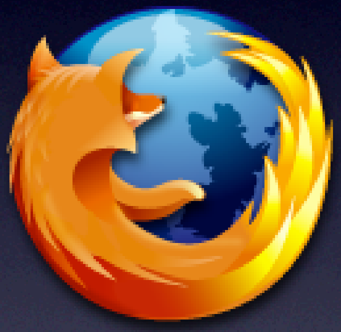
Our Assumptions

- Plaintext has some structure
- Plaintext is in a language we know

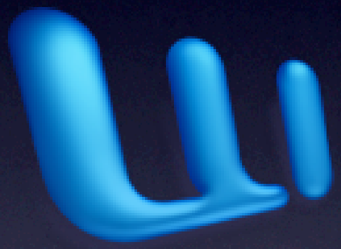
n-gram	count
⬤	2
a	2
p	2
l	1
e	2



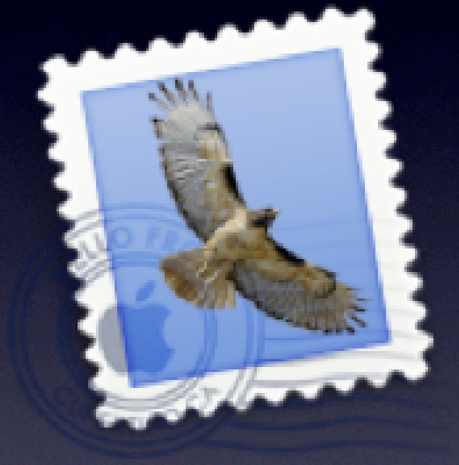
7 billion
characters

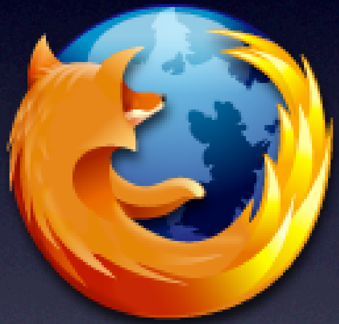


7 billion characters

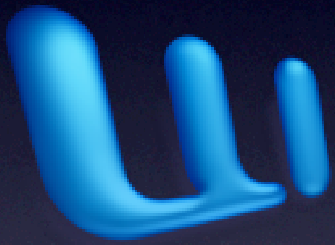


450 million characters

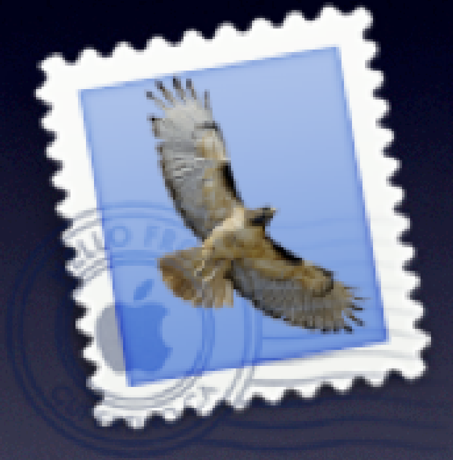




7 billion
characters

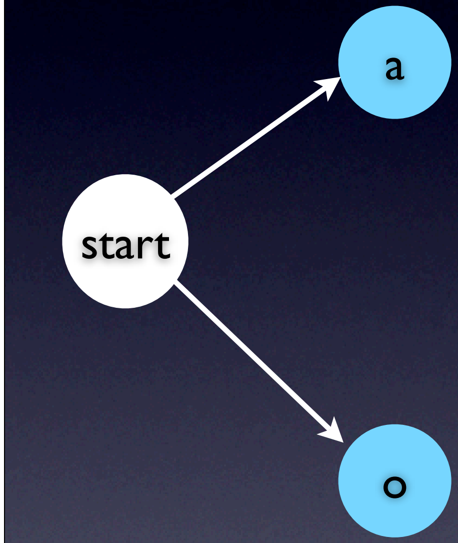


450 million
characters

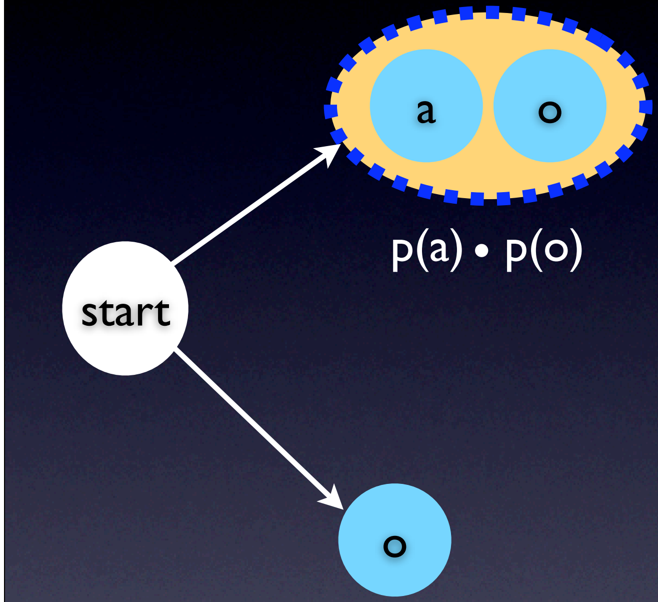


4 billion
characters

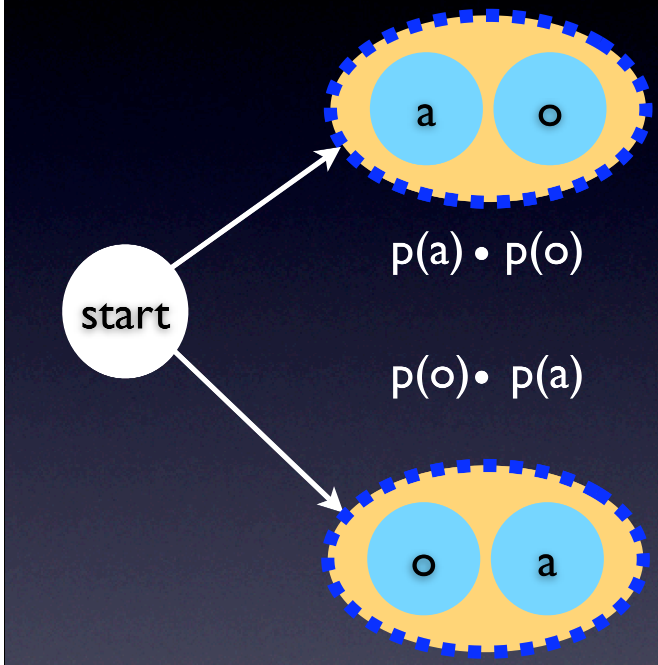
apple
orange



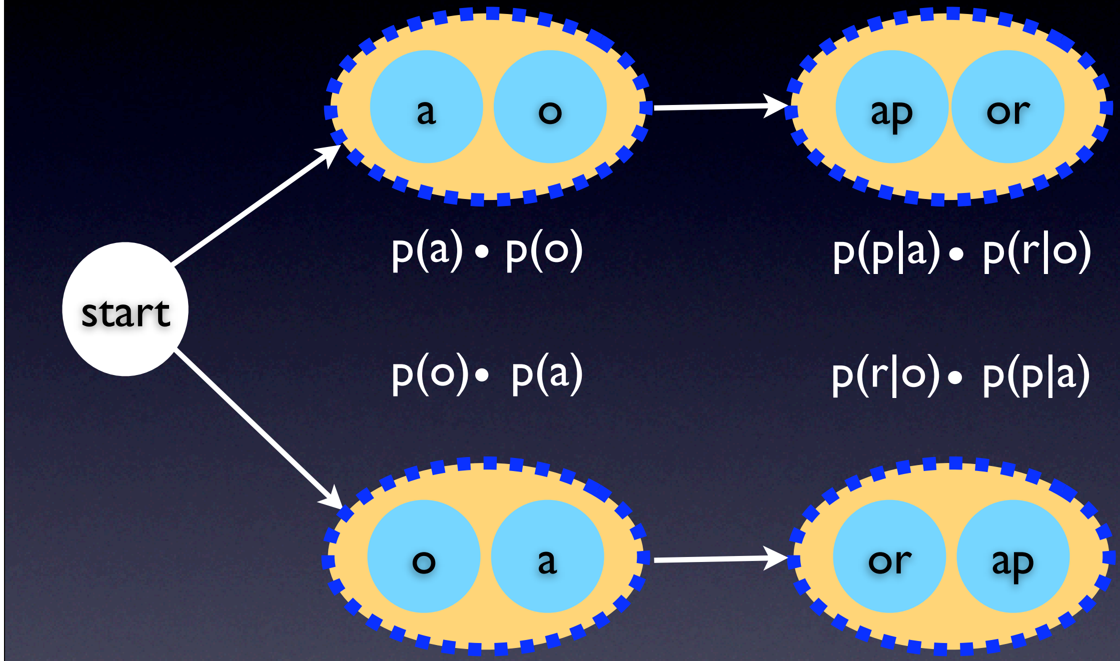
$P_0 \oplus P_1$	0e	02	11	02



$P_0 \oplus P_1$	0e	02	11	02



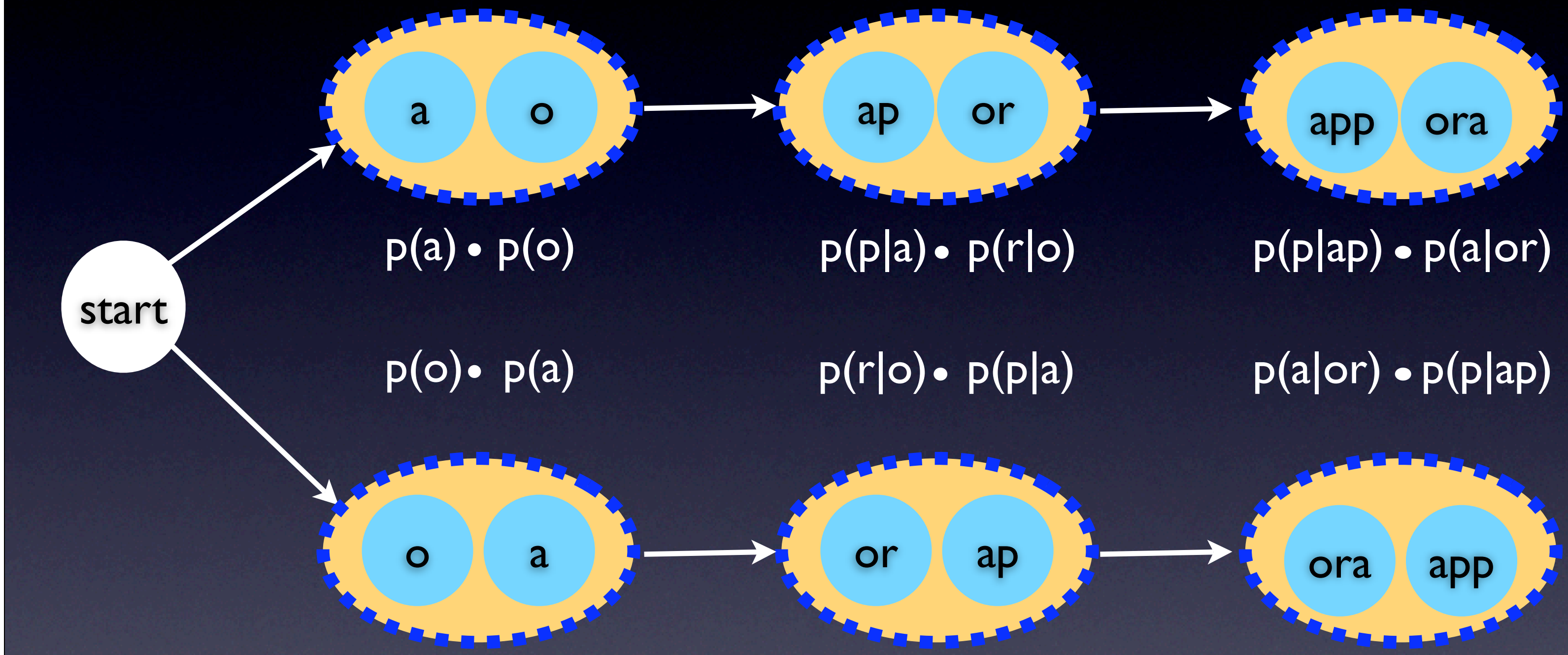
$P_0 \oplus P_1$	0e	02	11	02



$p(a) \bullet p(o)$ $p(p|a) \bullet p(r|o)$

$p(o) \bullet p(a)$ $p(r|o) \bullet p(p|a)$

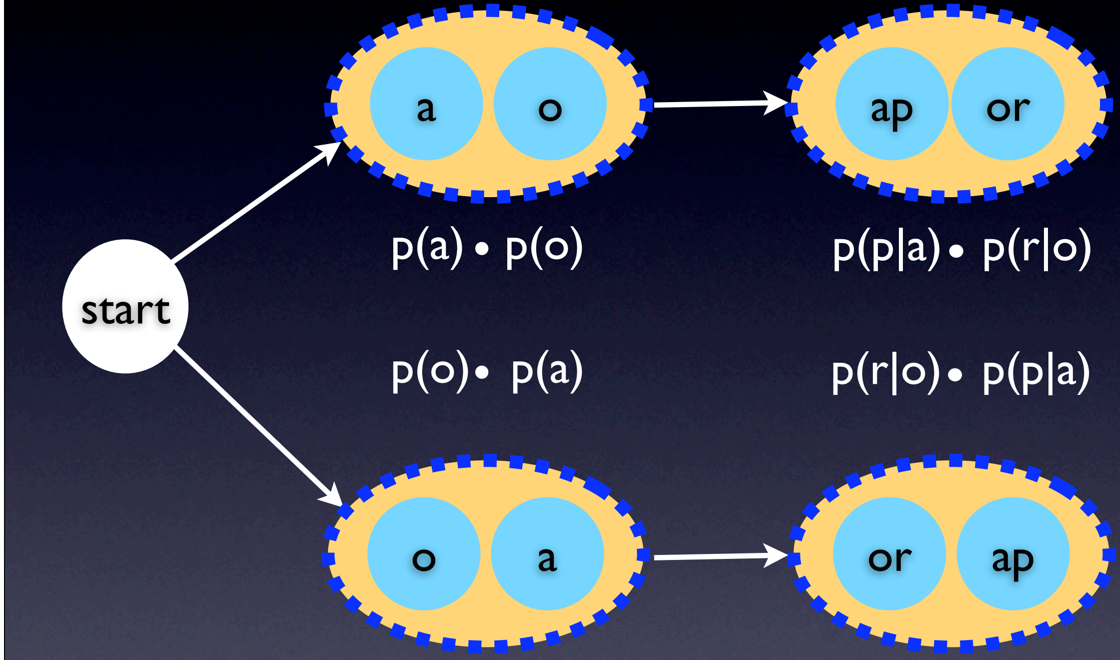
$P_0 \oplus P_1$	0e	02	11	02



$p(a) \bullet p(o)$ $p(p|a) \bullet p(r|o)$ $p(p|ap) \bullet p(a|or)$

$p(o) \bullet p(a)$ $p(r|o) \bullet p(p|a)$ $p(a|or) \bullet p(p|ap)$

$P_0 \oplus P_1$	0e	02	11	02



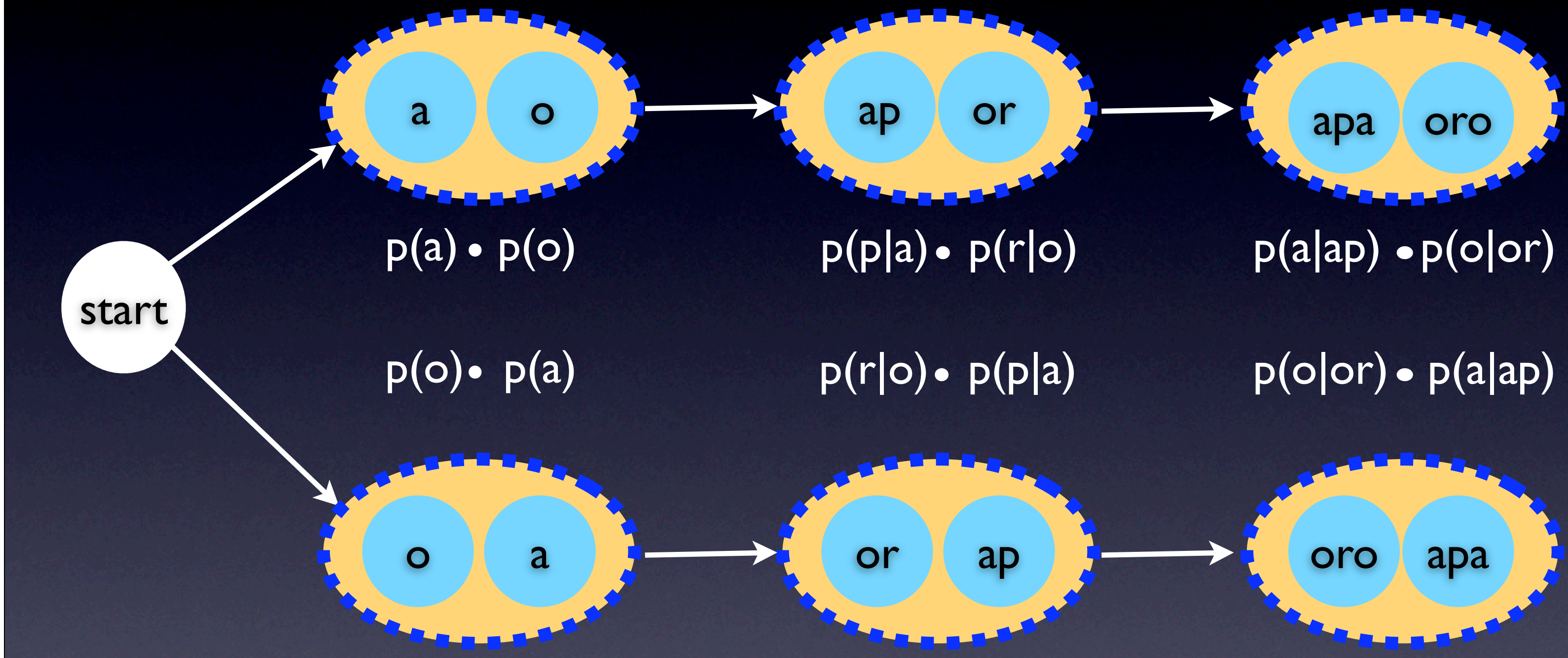
$p(a) \bullet p(o)$

$p(p|a) \bullet p(r|o)$

$p(o) \bullet p(a)$

$p(r|o) \bullet p(p|a)$

$P_0 \oplus P_1$	0e	02	0e	02



start

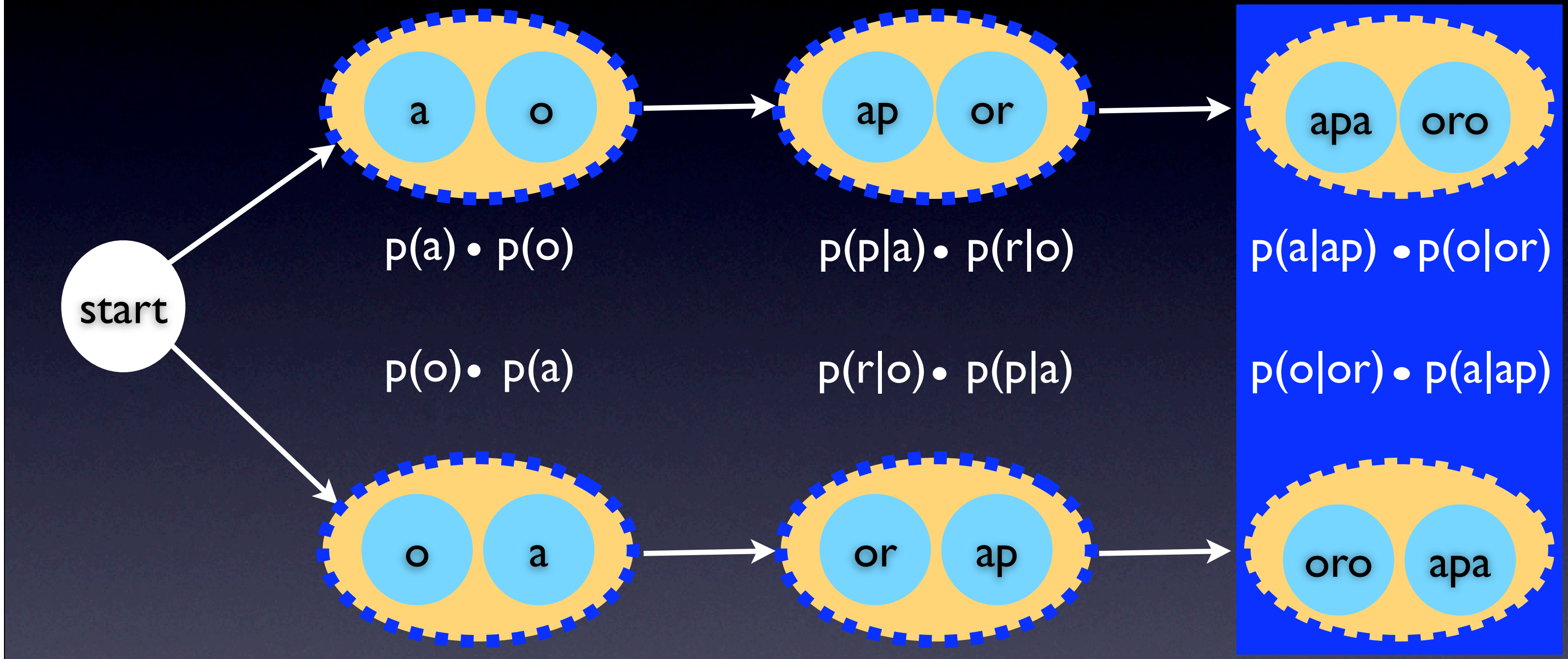
a o → ap or → apa oro

$p(a) \bullet p(o)$ $p(p|a) \bullet p(r|o)$ $p(a|ap) \bullet p(o|or)$

$p(o) \bullet p(a)$ $p(r|o) \bullet p(p|a)$ $p(o|or) \bullet p(a|ap)$

o a → or ap → oro apa

$P_0 \oplus P_1$	0e	02	0e	02



start

a o

ap or

apa oro

$p(a) \bullet p(o)$

$p(p|a) \bullet p(r|o)$

$p(a|ap) \bullet p(o|or)$

$p(o) \bullet p(a)$

$p(r|o) \bullet p(p|a)$

$p(o|or) \bullet p(a|ap)$

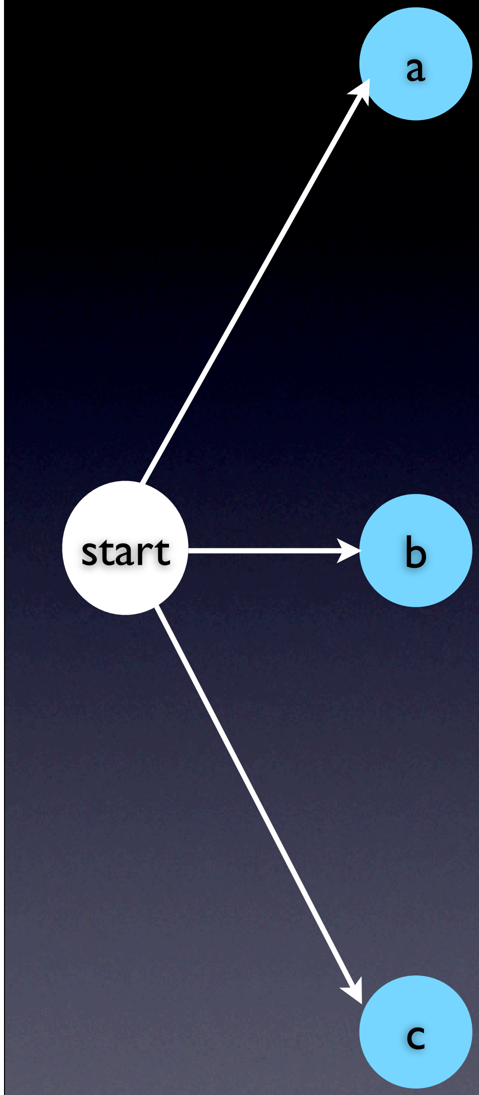
o a

or ap

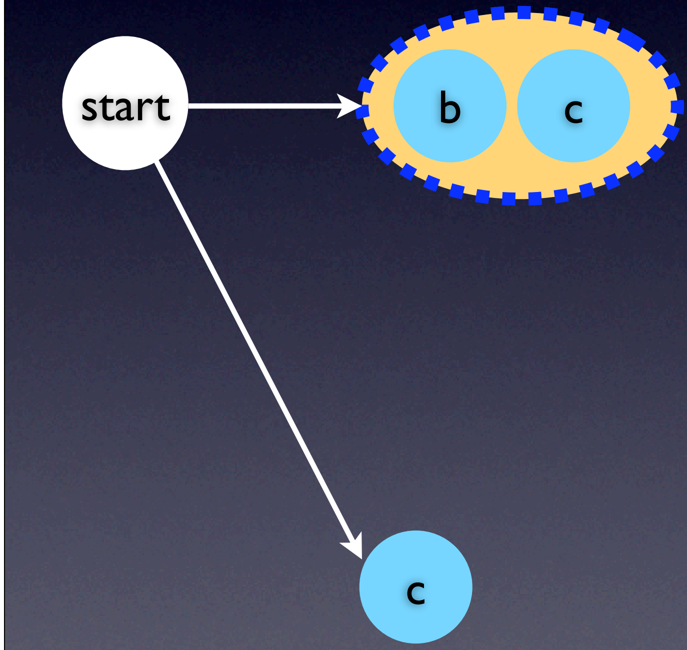
oro apa

$P_0 \oplus P_1$	0e	02	0e	02

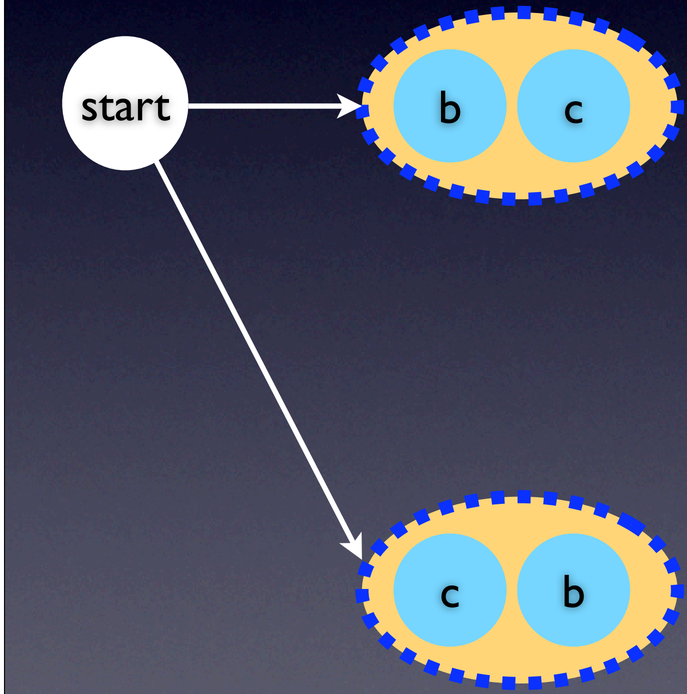
Memory/Computation



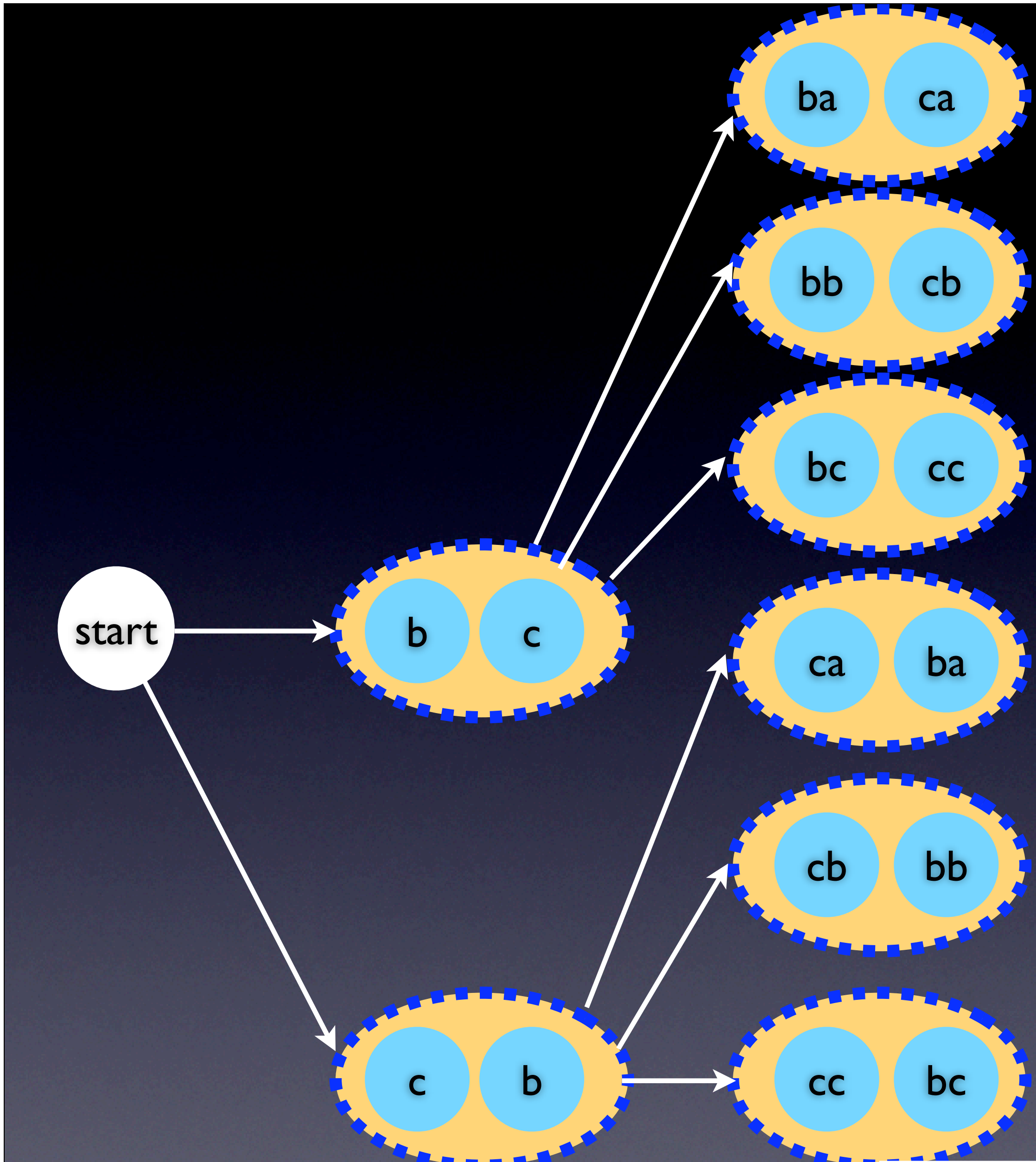
$P_2 \oplus P_3$	01	00	02	02



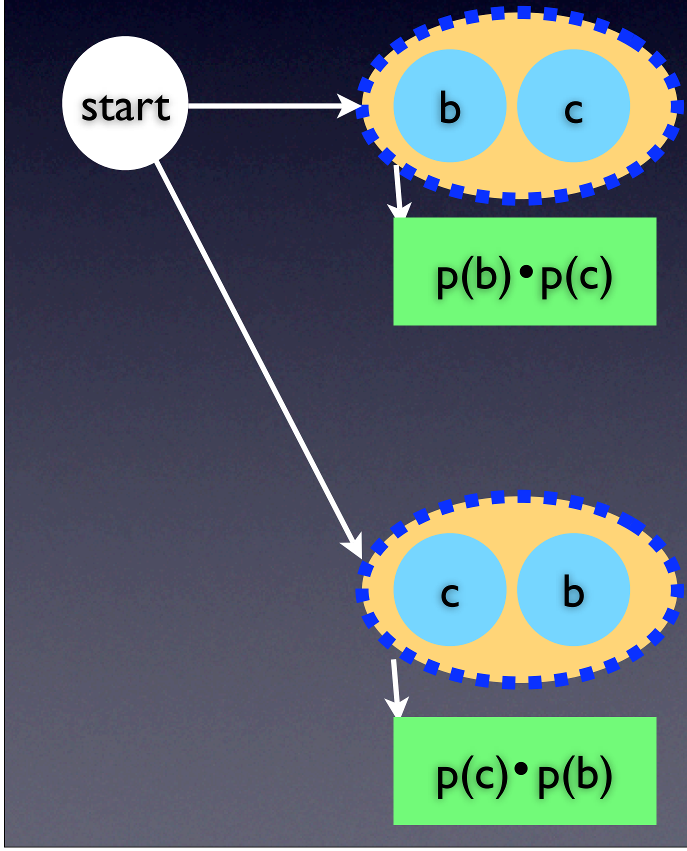
$P_2 \oplus P_3$	01	00	02	02



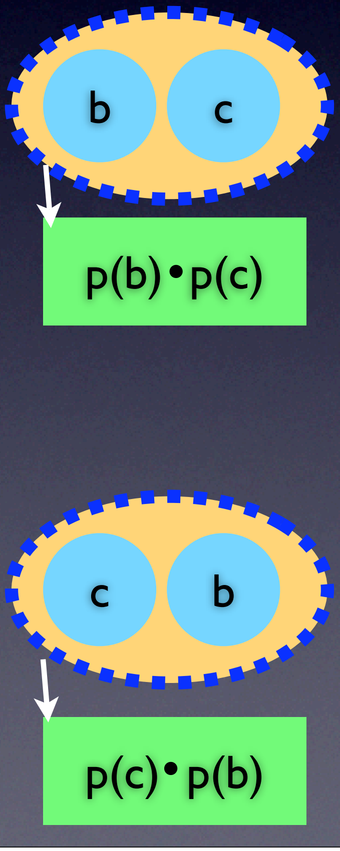
$P_2 \oplus P_3$	01	00	02	02



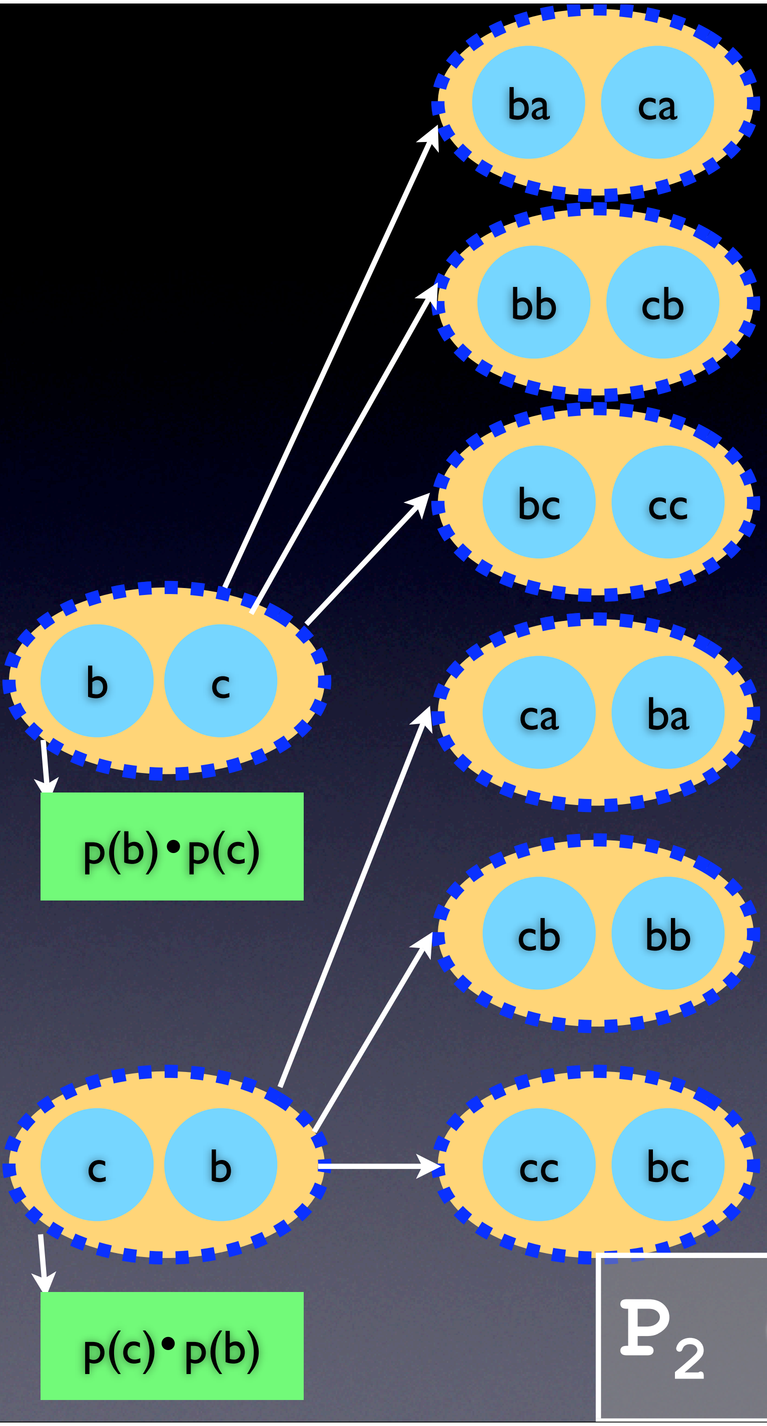
$P_2 \oplus P_3$	01	00	02	02



$P_2 \oplus P_3$	01	00	02	02



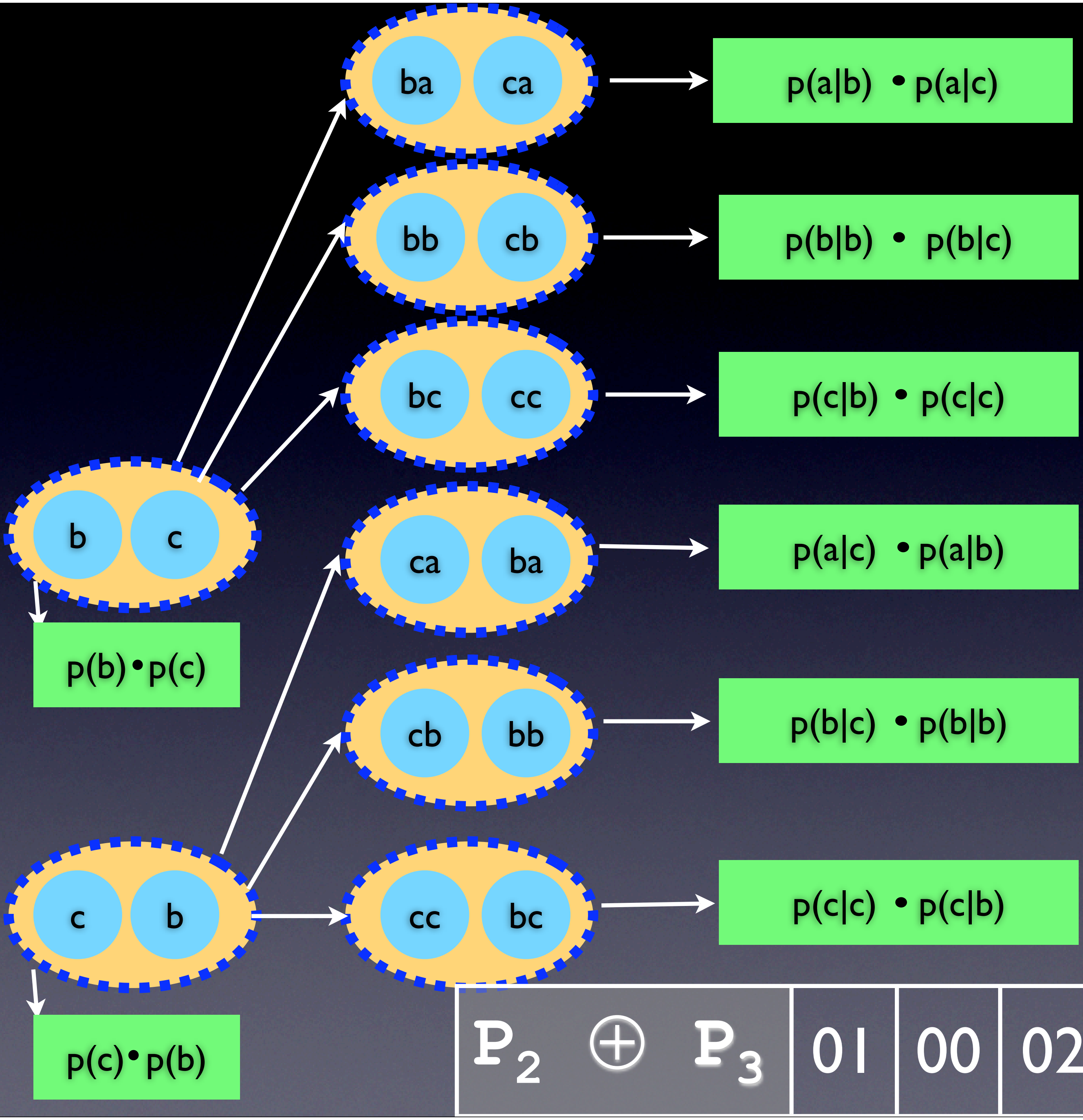
$P_2 \oplus P_3$	01	00	02	02



$p(b) \cdot p(c)$

$p(c) \cdot p(b)$

$P_2 \oplus P_3$	01	00	02	02



b c

$p(b) \cdot p(c)$

c b

$p(c) \cdot p(b)$

ba ca

$p(a|b) \cdot p(a|c)$

bb cb

$p(b|b) \cdot p(b|c)$

bc cc

$p(c|b) \cdot p(c|c)$

ca ba

$p(a|c) \cdot p(a|b)$

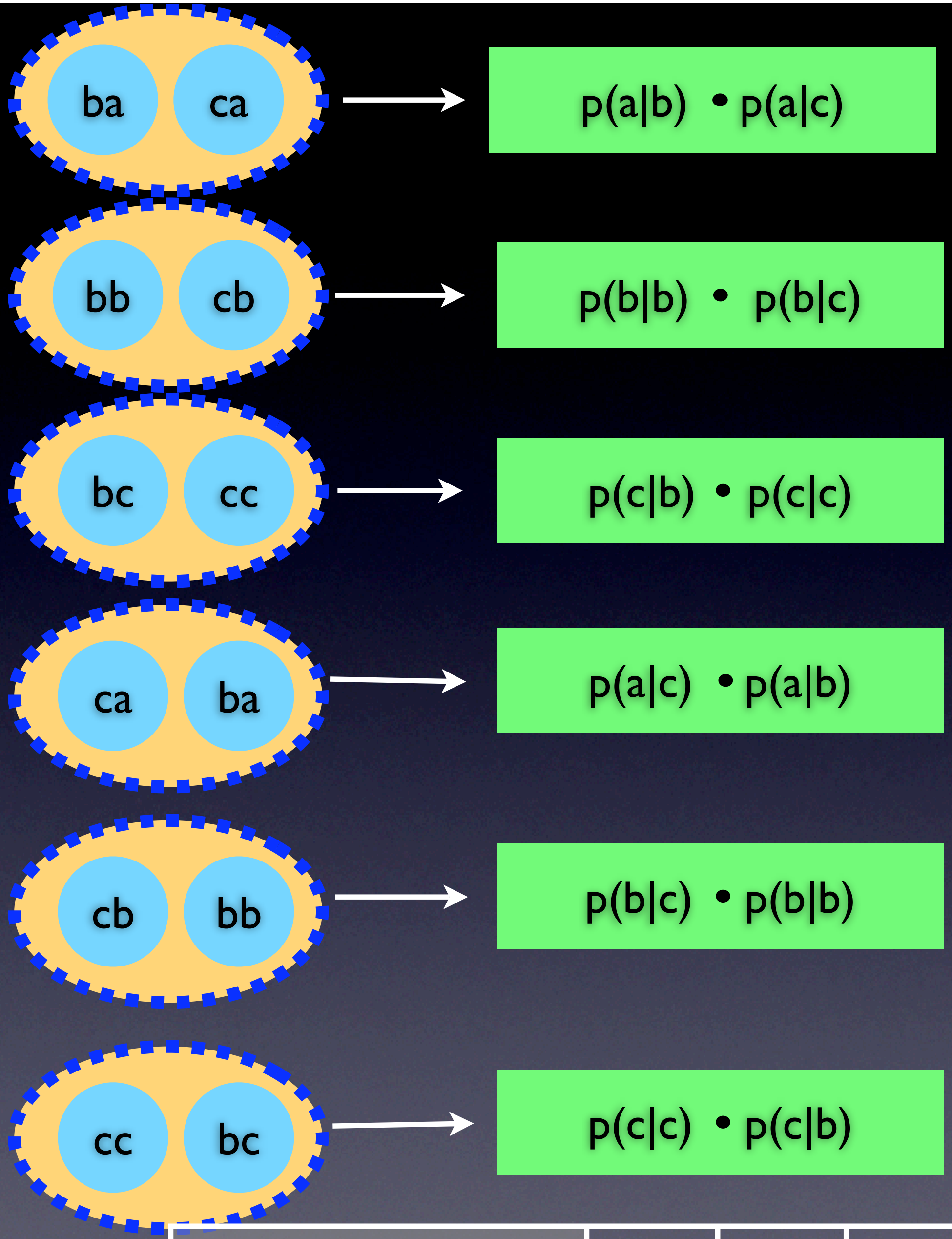
cb bb

$p(b|c) \cdot p(b|b)$

cc bc

$p(c|c) \cdot p(c|b)$

$P_2 \oplus P_3$	01	00	02	02



ba ca → $p(a|b) \bullet p(a|c)$

bb cb → $p(b|b) \bullet p(b|c)$

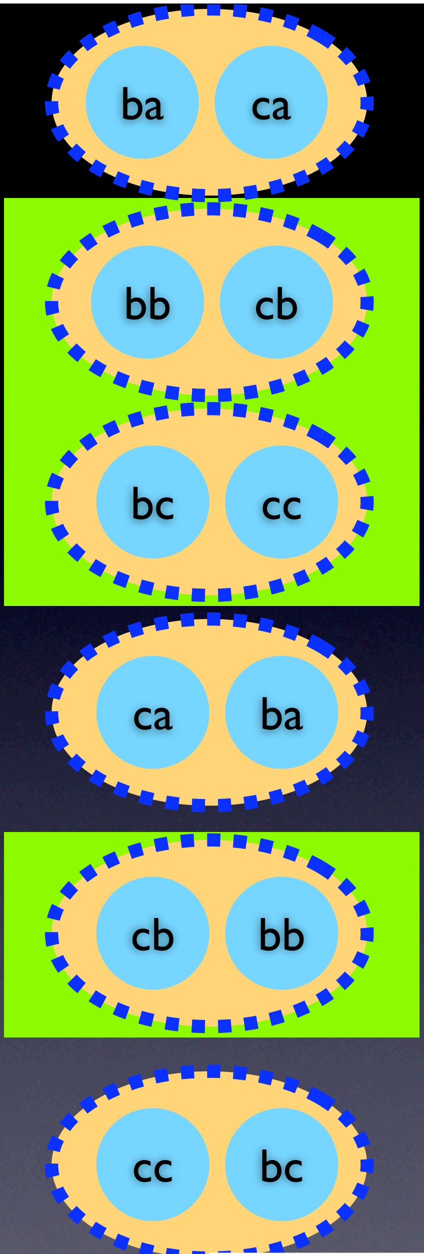
bc cc → $p(c|b) \bullet p(c|c)$

ca ba → $p(a|c) \bullet p(a|b)$

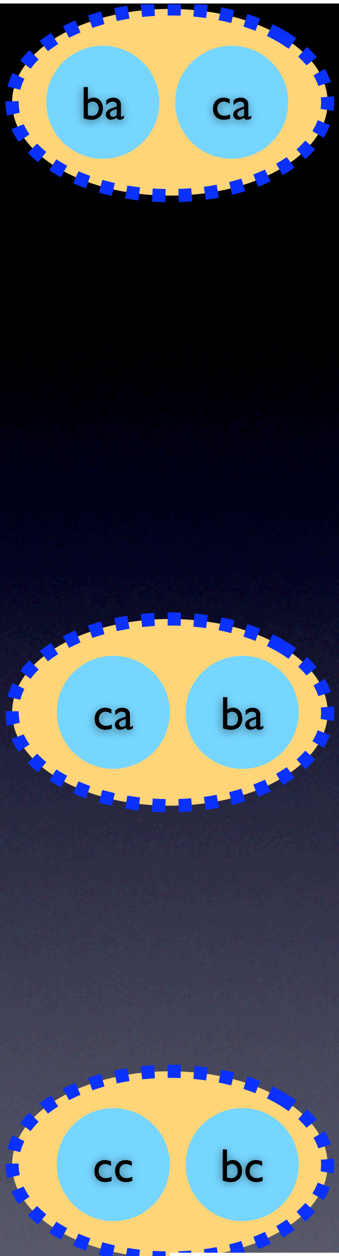
cb bb → $p(b|c) \bullet p(b|b)$

cc bc → $p(c|c) \bullet p(c|b)$

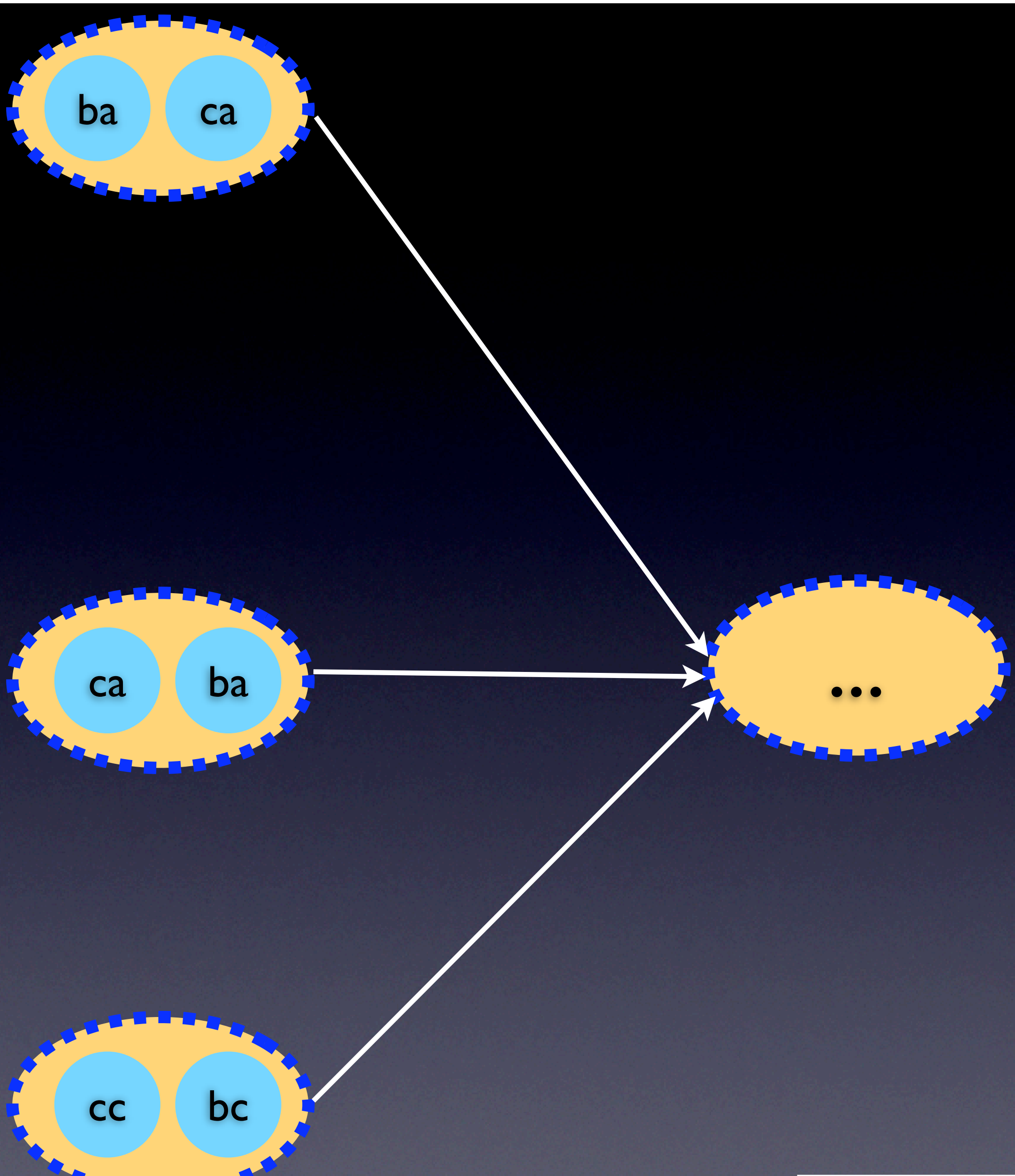
$P_2 \oplus P_3$	01	00	02	02



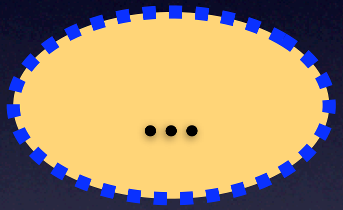
$P_2 \oplus P_3$	01	00	02	02



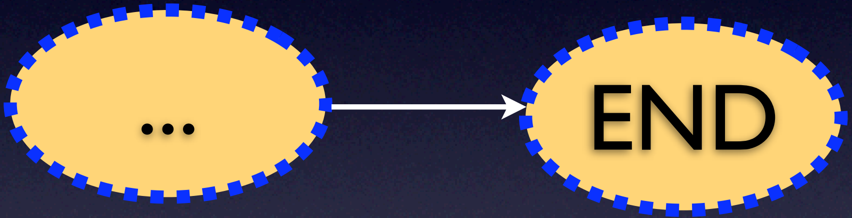
$P_2 \oplus P_3$	01	00	02	02



$P_2 \oplus P_3$	01	00	02	02



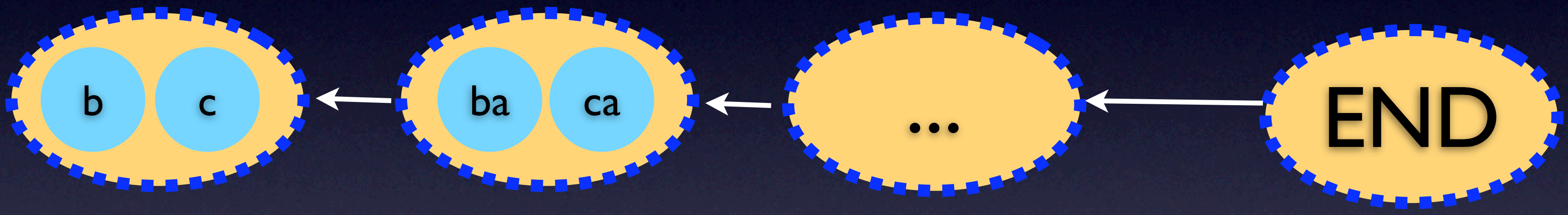
$P_2 \oplus P_3$	01	00	02	02



$P_2 \oplus P_3$	01	00	02	02

END

$P_2 \oplus P_3$	01	00	02	02



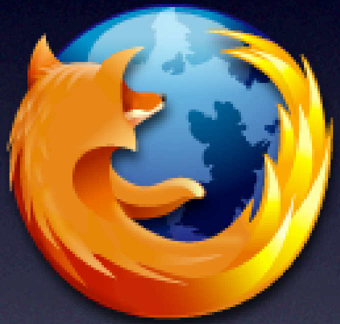
$P_2 \oplus P_3$	01	00	02	02

Commodity Hardware

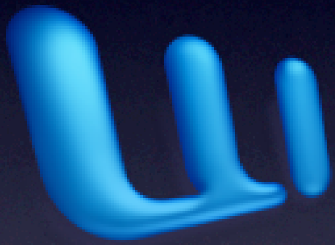
System	Dual Core Pentium 3 GHz
Memory	8 GB
Storage	1.2 TB

Model Build Time	~12 hours
Runtime	200 ms per byte
Memory Usage	~2 GB

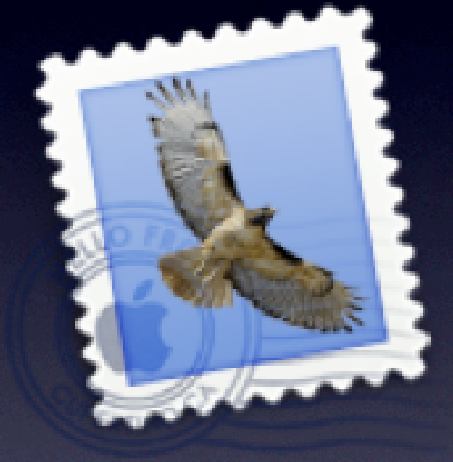
Our testing methodology



402,590 Files



98,699 Files



520,931 Files

402,590 Files	98,699 Files	520,931 Files
2,590 Files	8,699 Files	20,931 Files

402,590 Files	98,699 Files	520,931 Files
2,590 Files	8,699 Files	20,931 Files
50 Files	50 Files	50 Files

	Small		
HTML	90.64%		
E-mail	82.29%		
Documents	53.84%		

	Small	Medium	
HTML	90.64%	92.78%	
E-mail	82.29%	89.04%	
Documents	53.84%	53.05%	

	Small	Medium	Large
HTML	90.64%	92.78%	93.79%
E-mail	82.29%	89.04%	90.85%
Documents	53.84%	53.05%	52.72%

The Switching Problem

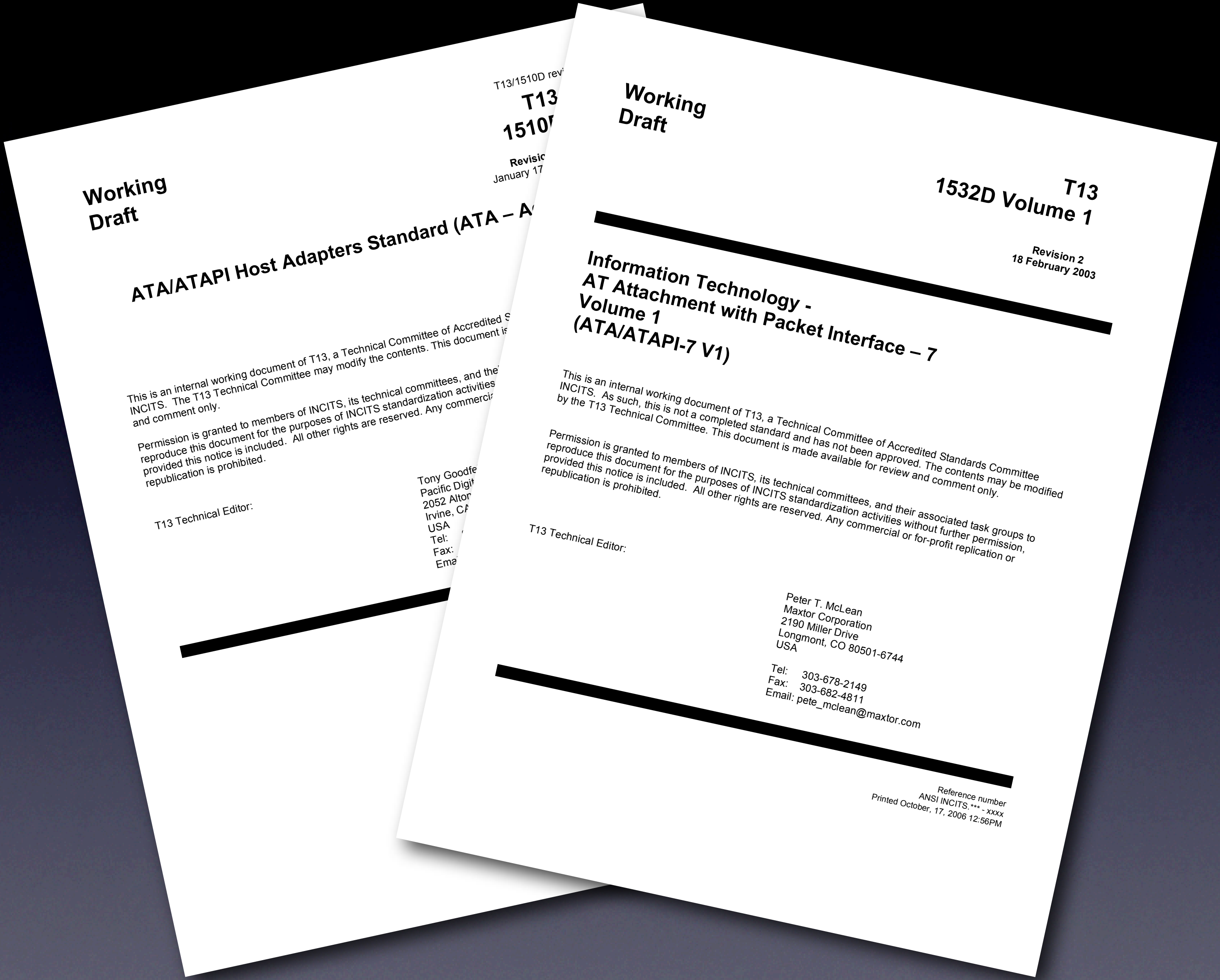
I want to remind you about our All-Employee Meeting this Tuesday, Oct. 23, at 10 a.m. Houston time at the Hyatt Regency. We obviously have a lot to talk about. Last week

Well I hope you have Dad doing some of the cleaning! You know how he always has an opinion but yet no participation. Anyway I hope you're doing fine. I'm fine

I want to remind you about our All-Employee Meeting this Tuesday, Oct. 23, at 10 a.m. Houston time at the Hyatt Regency participation. Anyway I hope you're doing fine. I'm fine and about to

Well I hope you have Dad doing some of the cleaning! You know how he always has an opinion but yet no. We obviously have a lot to talk about. Last week we reported third quarter earnings. We

Wu showed Word 2002 re-uses one time pad



T13/1510D revi

T13
1510D

Revisio
January 17

Working
Draft

ATA/ATAPI Host Adapters Standard (ATA – A

This is an internal working document of T13, a Technical Committee of Accredited S
INCITS. The T13 Technical Committee may modify the contents. This document is
and comment only.

Permission is granted to members of INCITS, its technical committees, and the
reproduce this document for the purposes of INCITS standardization activities
provided this notice is included. All other rights are reserved. Any commercia
republication is prohibited.

T13 Technical Editor:

Tony Goodfe
Pacific Digit
2052 Alton
Irvine, CA
USA
Tel:
Fax:
Ema

Working
Draft

T13
1532D Volume 1

Revision 2
18 February 2003

Information Technology -
AT Attachment with Packet Interface – 7
Volume 1
(ATA/ATAPI-7 V1)

This is an internal working document of T13, a Technical Committee of Accredited Standards Committee INCITS. As such, this is not a completed standard and has not been approved. The contents may be modified by the T13 Technical Committee. This document is made available for review and comment only.

Permission is granted to members of INCITS, its technical committees, and their associated task groups to reproduce this document for the purposes of INCITS standardization activities without further permission, provided this notice is included. All other rights are reserved. Any commercial or for-profit replication or republication is prohibited.

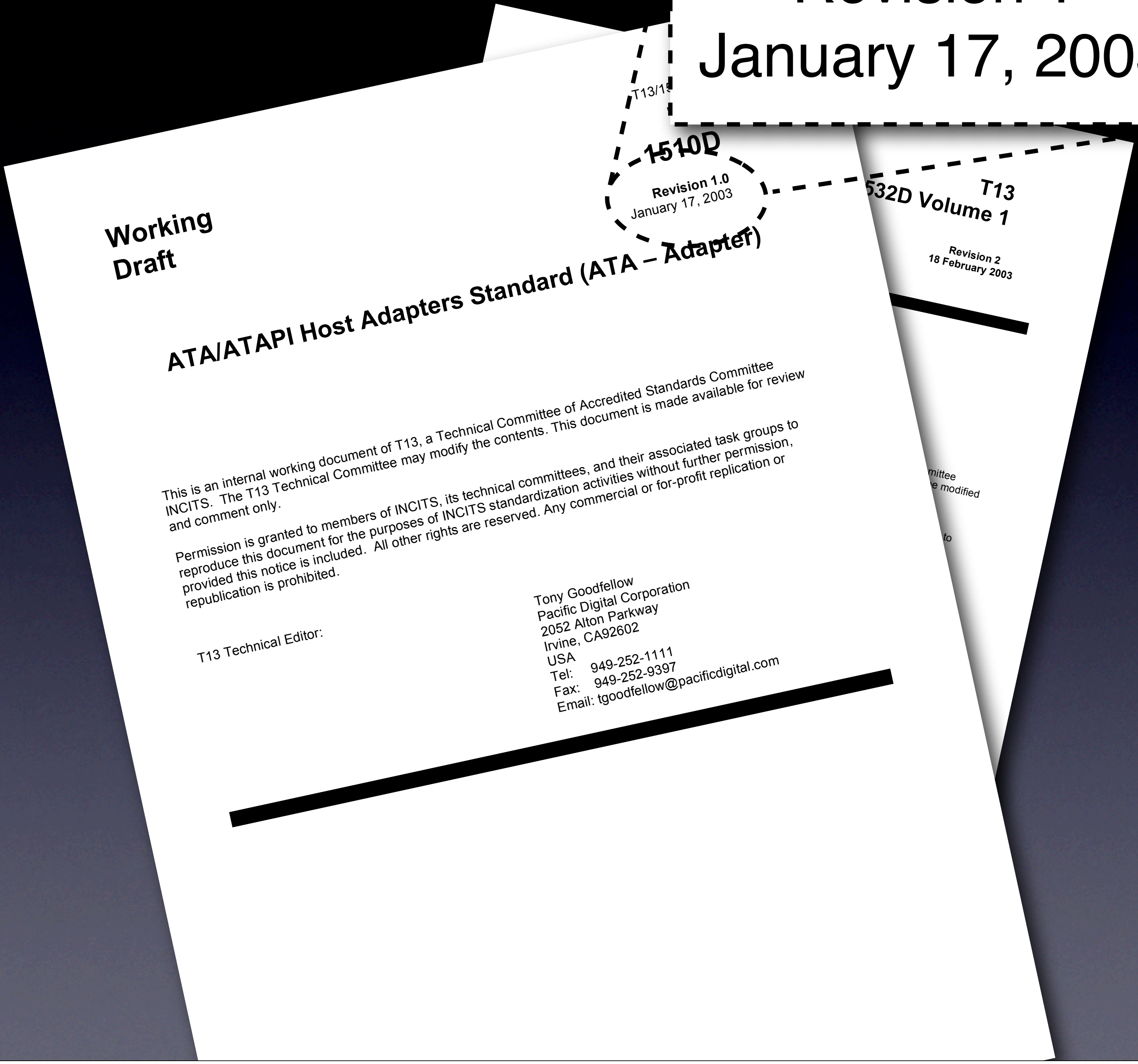
T13 Technical Editor:

Peter T. McLean
Maxtor Corporation
2190 Miller Drive
Longmont, CO 80501-6744
USA

Tel: 303-678-2149
Fax: 303-682-4811
Email: pete_mclean@maxtor.com

Reference number
ANSI INCITS.*** - xxxx
Printed October, 17, 2006 12:56PM

Revision 1
January 17, 2003



T13/15

1510D

Revision 1.0
January 17, 2003

Working
Draft

ATA/ATAPI Host Adapters Standard (ATA – Adapter)

This is an internal working document of T13, a Technical Committee of Accredited Standards Committee INCITS. The T13 Technical Committee may modify the contents. This document is made available for review and comment only.

Permission is granted to members of INCITS, its technical committees, and their associated task groups to reproduce this document for the purposes of INCITS standardization activities without further permission, provided this notice is included. All other rights are reserved. Any commercial or for-profit replication or republication is prohibited.

T13 Technical Editor:

Tony Goodfellow
Pacific Digital Corporation
2052 Alton Parkway
Irvine, CA92602
USA
Tel: 949-252-1111
Fax: 949-252-9397
Email: tgoodfellow@pacificdigital.com

T13

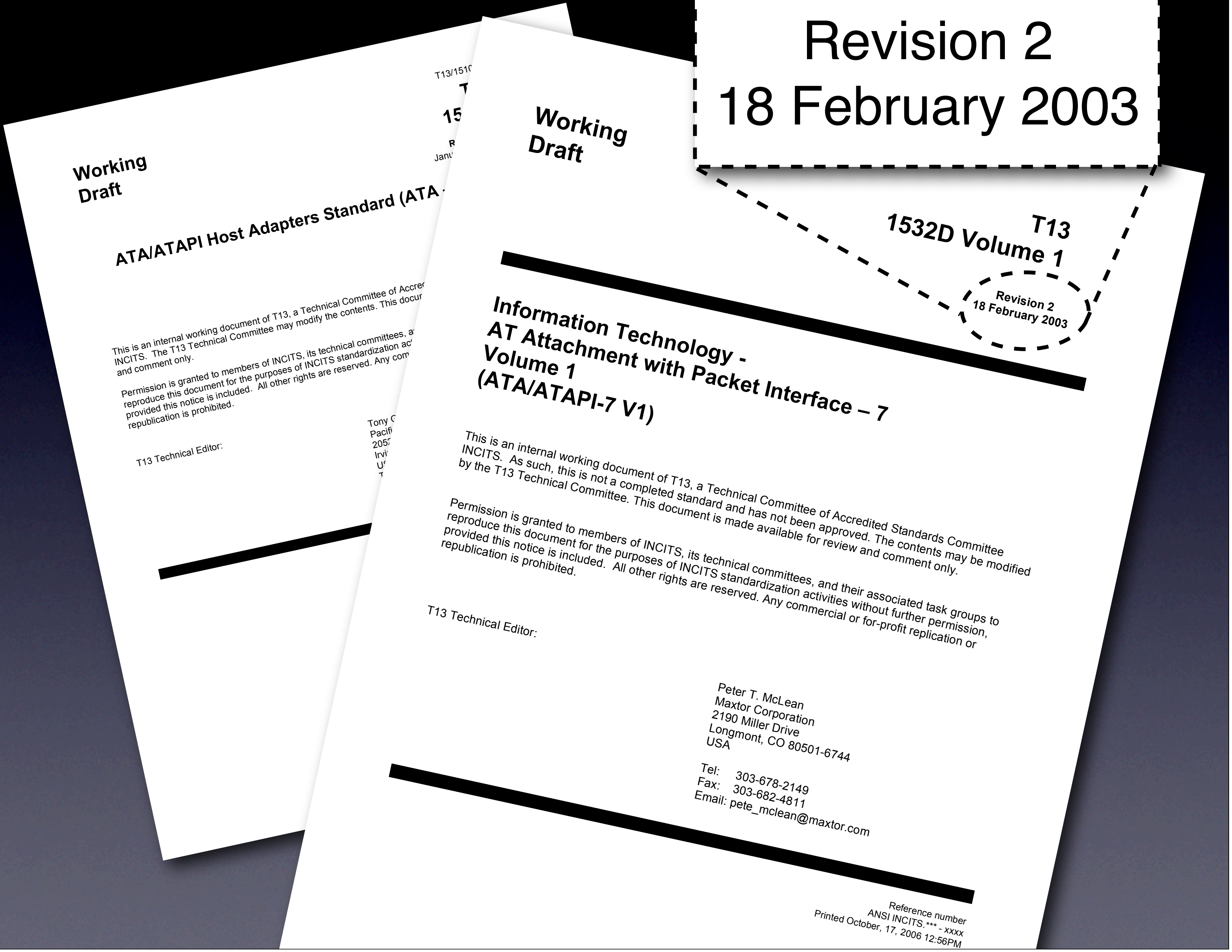
532D Volume 1

Revision 2
18 February 2003

mittee
e modified

to

Revision 2
18 February 2003



T13/1510
T
15
R
Janu

Working
Draft

ATA/ATAPI Host Adapters Standard (ATA -

This is an internal working document of T13, a Technical Committee of Accre
INCITS. The T13 Technical Committee may modify the contents. This docu
and comment only.

Permission is granted to members of INCITS, its technical committees, a
reproduce this document for the purposes of INCITS standardization ac
provided this notice is included. All other rights are reserved. Any com
republication is prohibited.

T13 Technical Editor:

Tony G
Pacif
2052
Irvi
U

Working
Draft

T13
1532D Volume 1

Revision 2
18 February 2003

Information Technology -
AT Attachment with Packet Interface – 7
Volume 1
(ATA/ATAPI-7 V1)

This is an internal working document of T13, a Technical Committee of Accredited Standards Committee INCITS. As such, this is not a completed standard and has not been approved. The contents may be modified by the T13 Technical Committee. This document is made available for review and comment only.

Permission is granted to members of INCITS, its technical committees, and their associated task groups to reproduce this document for the purposes of INCITS standardization activities without further permission, provided this notice is included. All other rights are reserved. Any commercial or for-profit replication or republication is prohibited.

T13 Technical Editor:

Peter T. McLean
Maxtor Corporation
2190 Miller Drive
Longmont, CO 80501-6744
USA

Tel: 303-678-2149
Fax: 303-682-4811
Email: pete_mclean@maxtor.com

Reference number
ANSI INCITS.*** - xxxx
Printed October, 17, 2006 12:56PM

November 13, 2002 ATA/ATAPI Host Adapters Standard (ATA   Adapter) This is an internal working document of T13, a Technical Committee of Accredited Standards Committee INCITS. The T13 Technical Committee may modify the contents. This document is made available for review and comment only. Permission is granted to members of INCITS, its technical committees, and their associated task groups to reproduce

November 13, 2002 ATA/ATAPI Host Adapters Standard (ATF; h Packet) This is no internal working document of T13, a Technical Committee of Accredited Standards Committee INCITS. The T13 Technical Committee may modify the contents. This document is made available and has not been approved. The contents may be modified by the T13 Technical technical committees, and their associated task groups to reproduce

November 13, 2002 ATA/ATAPI Host Adapters Standard (ATF; h Packet) This is no internal working document of T13, a Technical Committee of Accredited Standards Committee INCITS. The T13 Technical Committee may modify the contents. This document is made available and has not been approved. The contents may be modified by the T13 Technical technical committees, and their associated task groups to reproduce

	Exact	Pairwise
HTML	93.79%	99.45%
E-mail	90.85%	98.41%
Documents	52.72%	75.91%

Attack at Dawn $\oplus$ doQvYcSWIPyXaC

Take the Beach $\oplus$ doQvYcSWIPyXaC

Attack at Dawn ⊕ doQvYcSWIPyXaC

Take the Beach ⊕ doQvYcSWIPyXaC

Bring me Cakes ⊕ doQvYcSWIPyXaC

Attack at Dawn ⊕ doQvYcSWIPyXaC

Take the Beach ⊕ doQvYcSWIPyXaC

⊕

Attack at Dawn

⊕

Take the Beach

Attack at Dawn ⊕ doQvYcSWIPyXaC

⊕ Bring me Cakes ⊕ doQvYcSWIPyXaC

Attack at Dawn

⊕

Bring me Cakes

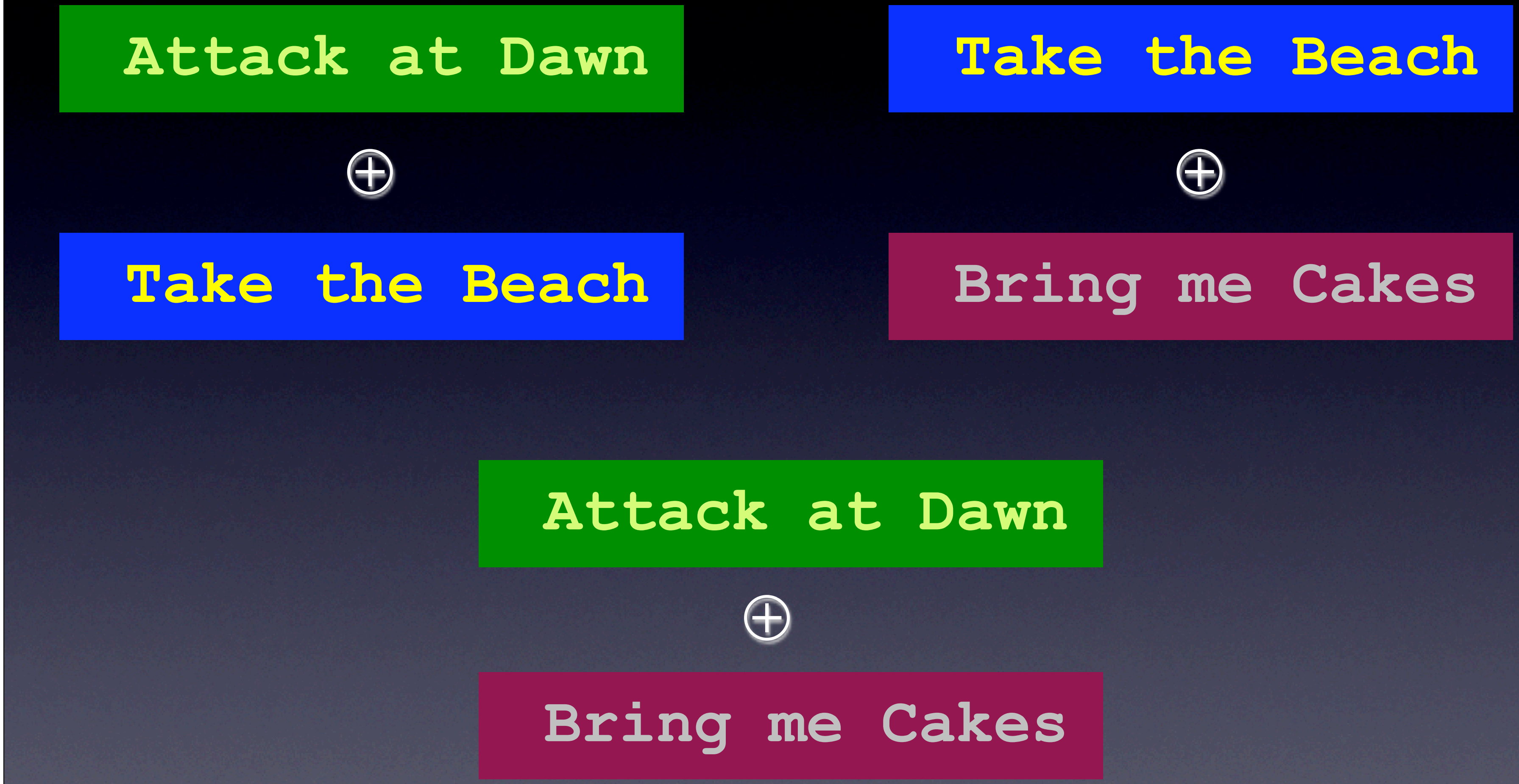
Take the Beach ⊕ doQvYcSWIPyXaC

⊕ Bring me Cakes ⊕ doQvYcSWIPyXaC

Take the Beach

⊕

Bring me Cakes



Attack at Dawn

⊕

Take the Beach

Take the Beach

⊕

Bring me Cakes

Attack at Dawn

⊕

Bring me Cakes

Attack at Dawn

$\oplus$

Take the Beach

Attack at Dawn

⊕

Take the Beach

Attack at Dawn

⊕

Take the Beach

Take the Beach

⊕

Bring me Cakes

	Small		
HTML	99.96%		
E-mail	98.24%		
Documents	69.92%		

	Small	Medium	
HTML	99.96%	99.95%	
E-mail	98.24%	98.33%	
Documents	69.92%	71.11%	

	Small	Medium	Large
HTML	99.96%	99.95%	99.95%
E-mail	98.24%	98.33%	98.34%
Documents	69.92%	71.11%	69.39%

⊕

	Large
HTML	93.79%
E-mail ⊕ HTML	96.60%
E-mail	90.85%

Conclusions

Able to recover plaintext with over 99% accuracy

Conclusions

Able to recover plaintext with over 99% accuracy

Technique works on different document types

Conclusions

Able to recover plaintext with over 99% accuracy

Technique works on different document types

Keystream reuse is a real problem