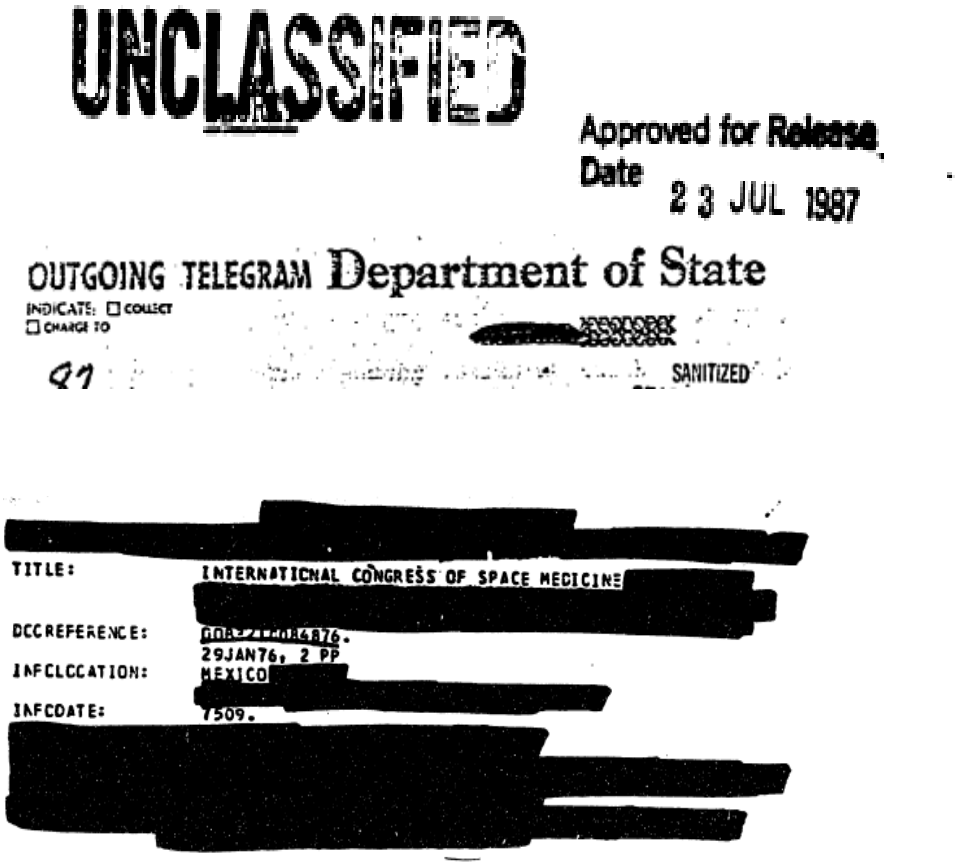
Textual Analysis of Government Declassification Patterns

Hanna Wallach

wallach@cs.umass.edu
http://www.cs.umass.edu/~wallach/

Transparency in the US

[ISOO, 2011]


UNCLASSIFIED

Approved for Release
Date 23 JUL 1987

OUTGOING TELEGRAM Department of State

INDICATE: ☐ COLLECT
☐ CHARGE TO

SANITIZED

TITLE: INTERNATIONAL CONGRESS OF SPACE MEDICINE

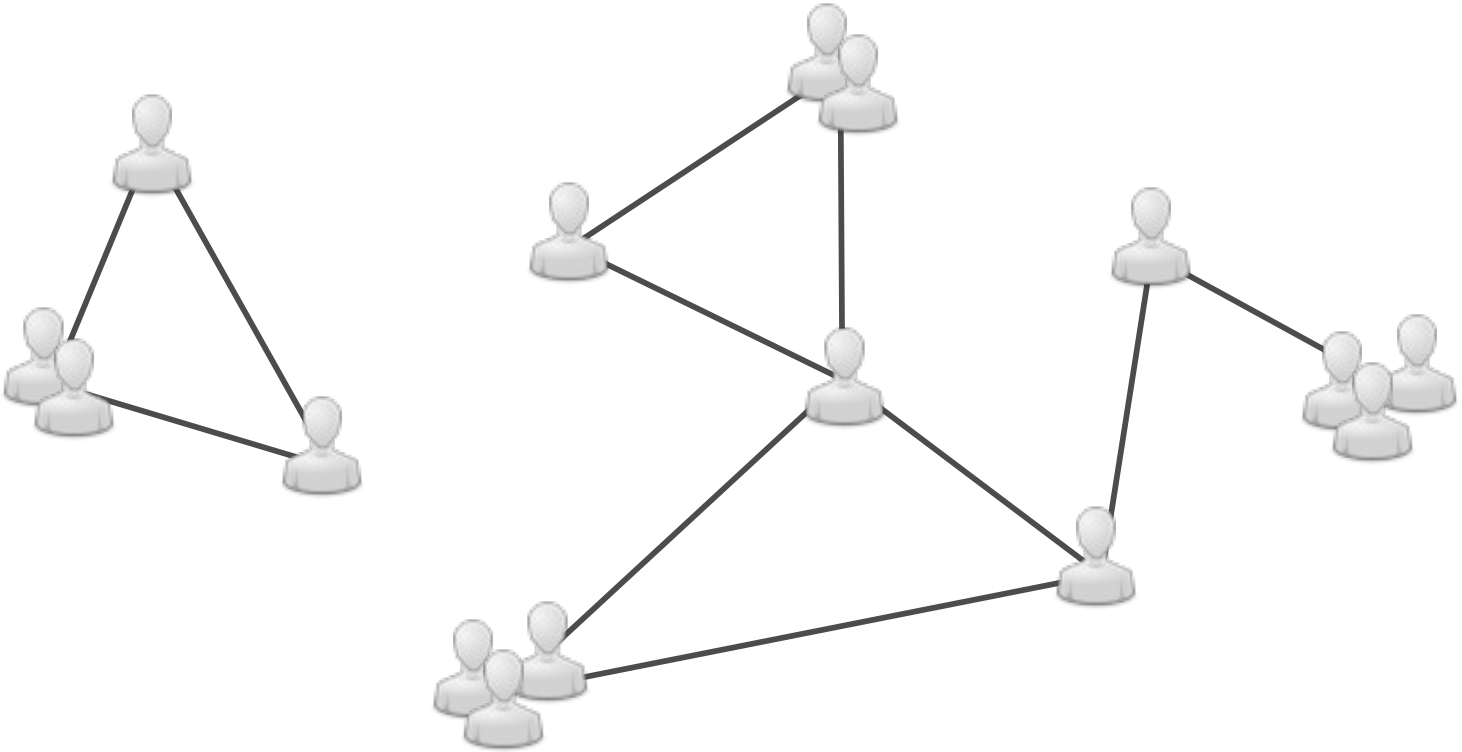
DOCREFERENCE: 29JAN76, 2 PP

INFOLOCATION: MEXICO

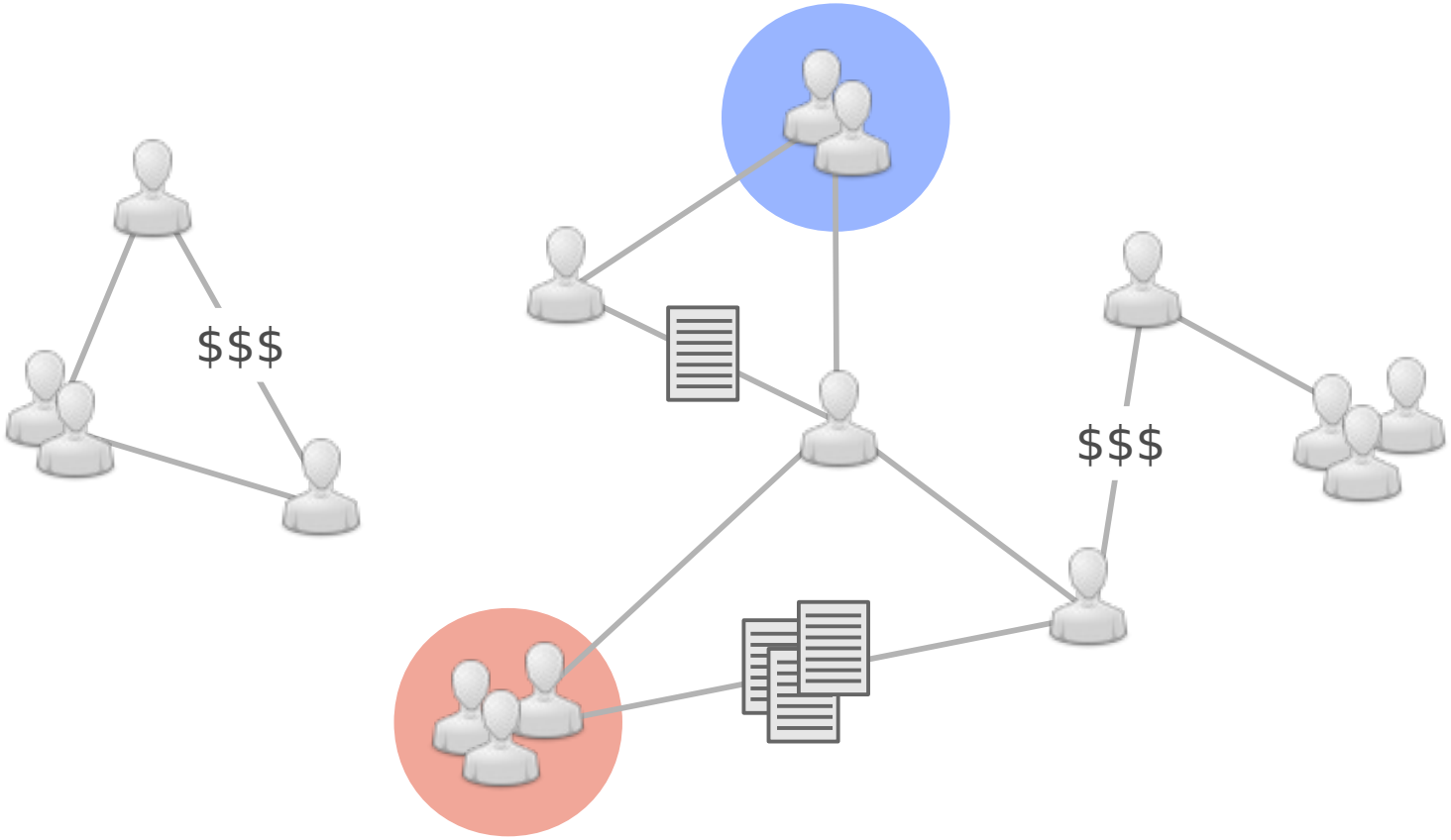
INFODATE: 7509.

- 52.8 million pages reviewed for declassification
- 26.7 million pages declassified
- $11.36 billion spent on administration of the US government classification system

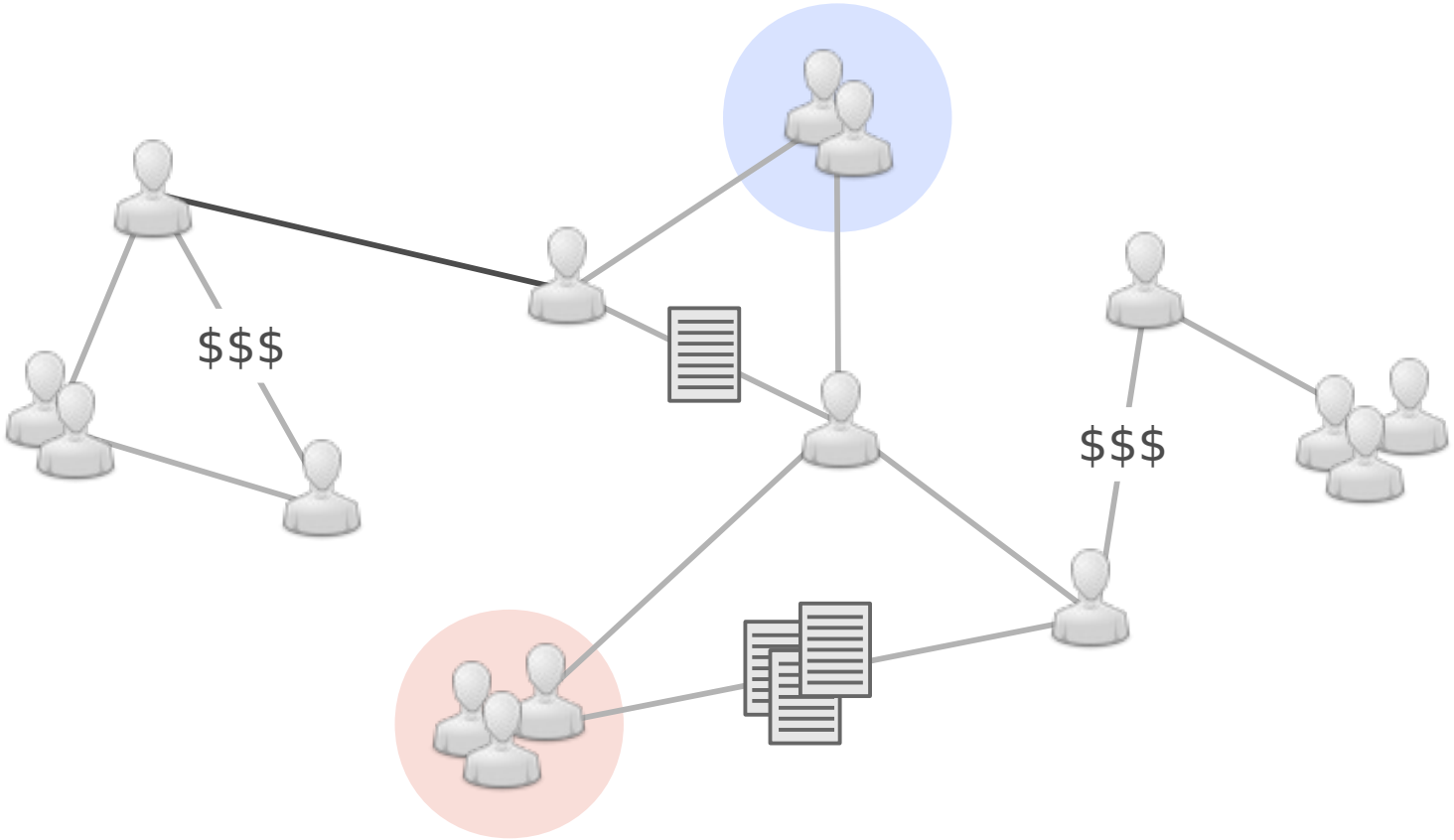
Complex Social Processes



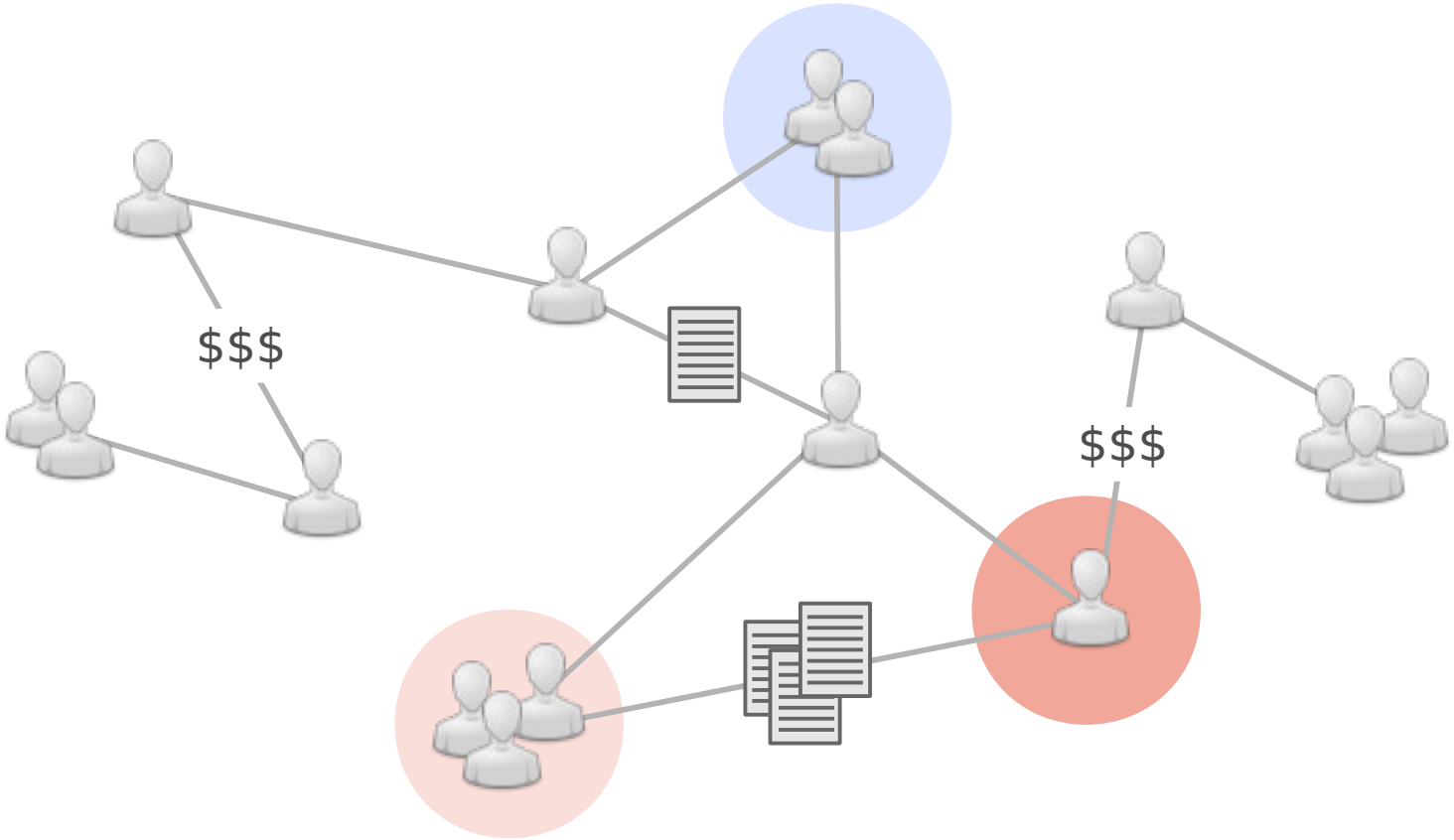
Complex Social Processes



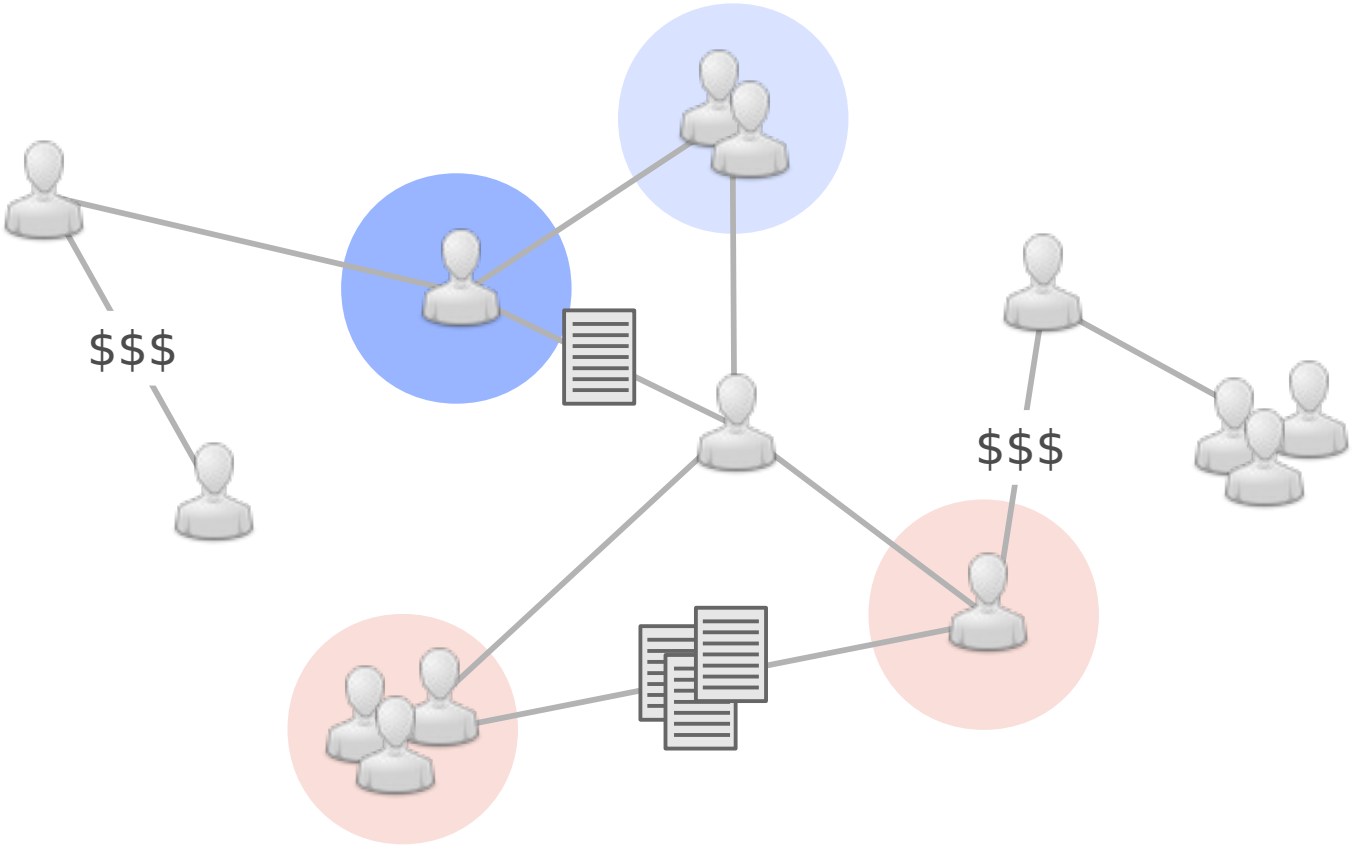
Complex Social Processes



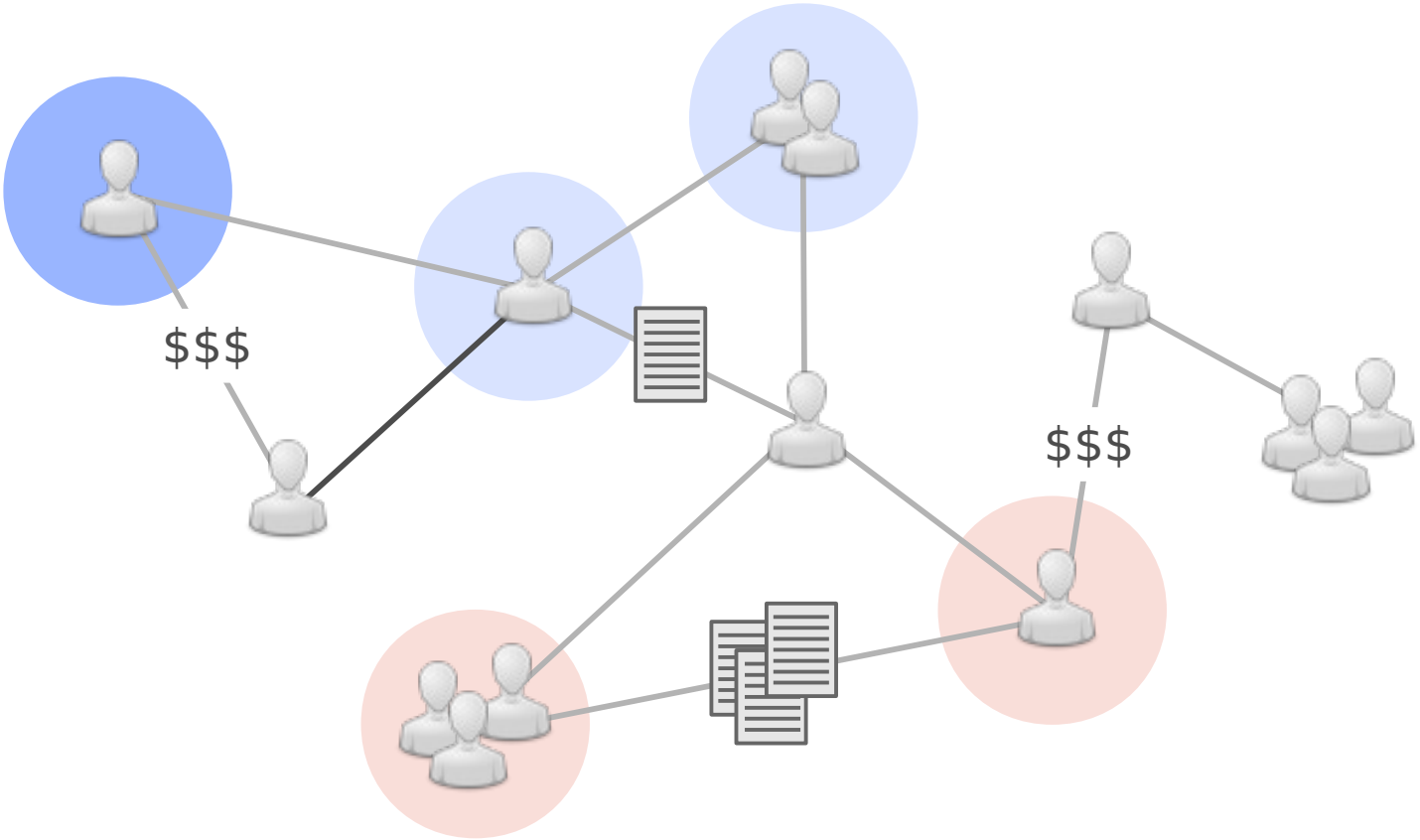
Complex Social Processes



Complex Social Processes



Complex Social Processes



Declassified Documents

[Gale, 2012]

~~SECRET~~ NO FOREIGN DISSEM

CENTRAL INTELLIGENCE AGENCY
WASHINGTON, D.C. 20505

29 January 1968

MEMORANDUM FOR: The Honorable Walt W. Rostow
Special Assistant to the President
The White House

SUBJECT : Coal and Electric Power Shortages in Communist China

1. Al Jenkins asked that we prepare the attached memorandum on shortages of coal and electric power in Communist China for your information. We have also included excerpts from individual reports of shortages to give you some feeling for the information available.

2. While there is no question that the shortages are widespread, it is extremely difficult to quantify the decline in industrial output caused by these shortages or by other effects of the Cultural Revolution.

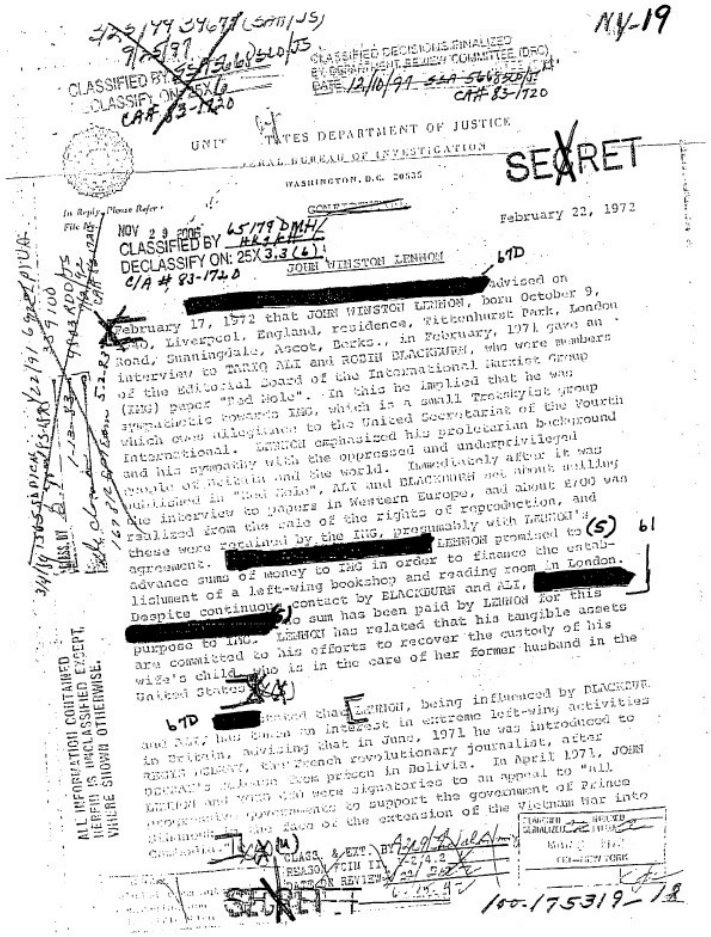
Edward W. Proctor
EDWARD W. PROCTOR
Acting Deputy Director for Intelligence

Attachment:
Subject Report

DECLASSIFIED
E.O. 12958, Sec. 3.6
NLJ 92-193
By Cb, NARA Date 10-31-97

- Date issued
- Date declassified
- Document type
- Source institution
- Classification level
- Document text

Text Tells All


NY-19

CLASSIFIED DECISIONS FINALIZED BY DECLASSIFICATION REVIEW COMMITTEE (DRC) DATE 12/10/91

CLASSIFIED BY

CA# 83-1720

UNITED STATES DEPARTMENT OF JUSTICE

FEDERAL BUREAU OF INVESTIGATION

WASHINGTON, D.C. 20535

SECRET

February 22, 1972

NOV 29 2006 CLASSIFIED BY 65179 DMH/ DECLASSIFY ON: 25X 3.3(b)(1) CIA # 83-1720

JOHN WINSTON LENNON

b7D

[redacted] advised on February 17, 1972 that JOHN WINSTON LENNON, born October 9, [illegible], Liverpool, England, residence, Tittenhurst Park, London Road, Sunningdale, Ascot, Berks., in February, 1971 gave an interview to TARIQ ALI and ROBIN BLACKBURN, who were members of the Editorial Board of the International Marxist Group (IMG) paper "Red Mole". In this he implied that he was sympathetic towards IMG, which is a small Trotskyist group which owes allegiance to the United Secretariat of the Fourth International. LENNON emphasized his proletarian background and his sympathy with the oppressed and underprivileged people of Britain and the world. Immediately after it was published in "Red Mole", ALI and BLACKBURN met about selling the interview to papers in Western Europe, and about £700 was realized from the sale of the rights of reproduction, and these were retained by the IMG, presumably with LENNON's agreement. [redacted] LENNON promised to advance sums of money to IMG in order to finance the establishment of a left-wing bookshop and reading room in London. Despite continuous [redacted] contact by BLACKBURN and ALI, [redacted] no sum has been paid by LENNON for this purpose to IMG. LENNON has related that his tangible assets are committed to his efforts to recover the custody of his wife's child, [redacted] who is in the care of her former husband in the United States. (5) 61

b7D [redacted] stated that LENNON, being influenced by BLACKBURN and ALI, has taken an interest in extreme left-wing activities in Britain, advising that in June, 1971 he was introduced to REGIS DEBRAY, the French revolutionary journalist, after DEBRAY's release from prison in Bolivia. In April 1971, JOHN LENNON and YOKO ONO were signatories to an appeal to "all progressive governments to support the government of Prince Sihanouk in the face of the extension of the Vietnam War into Cambodia." (4)

CLASS. & EXT. BY REASON FCIM II DATE OF REVIEW

SECRET

100-175319-18

ALL INFORMATION CONTAINED HEREIN IS UNCLASSIFIED EXCEPT WHERE SHOWN OTHERWISE.

"The FBI has released a new cache of material on John Lennon. The file contains little, if any, new information about Lennon, though it does present some bizarre details, like a description of an antiwar activist trying to train a parrot to speak profanities.."

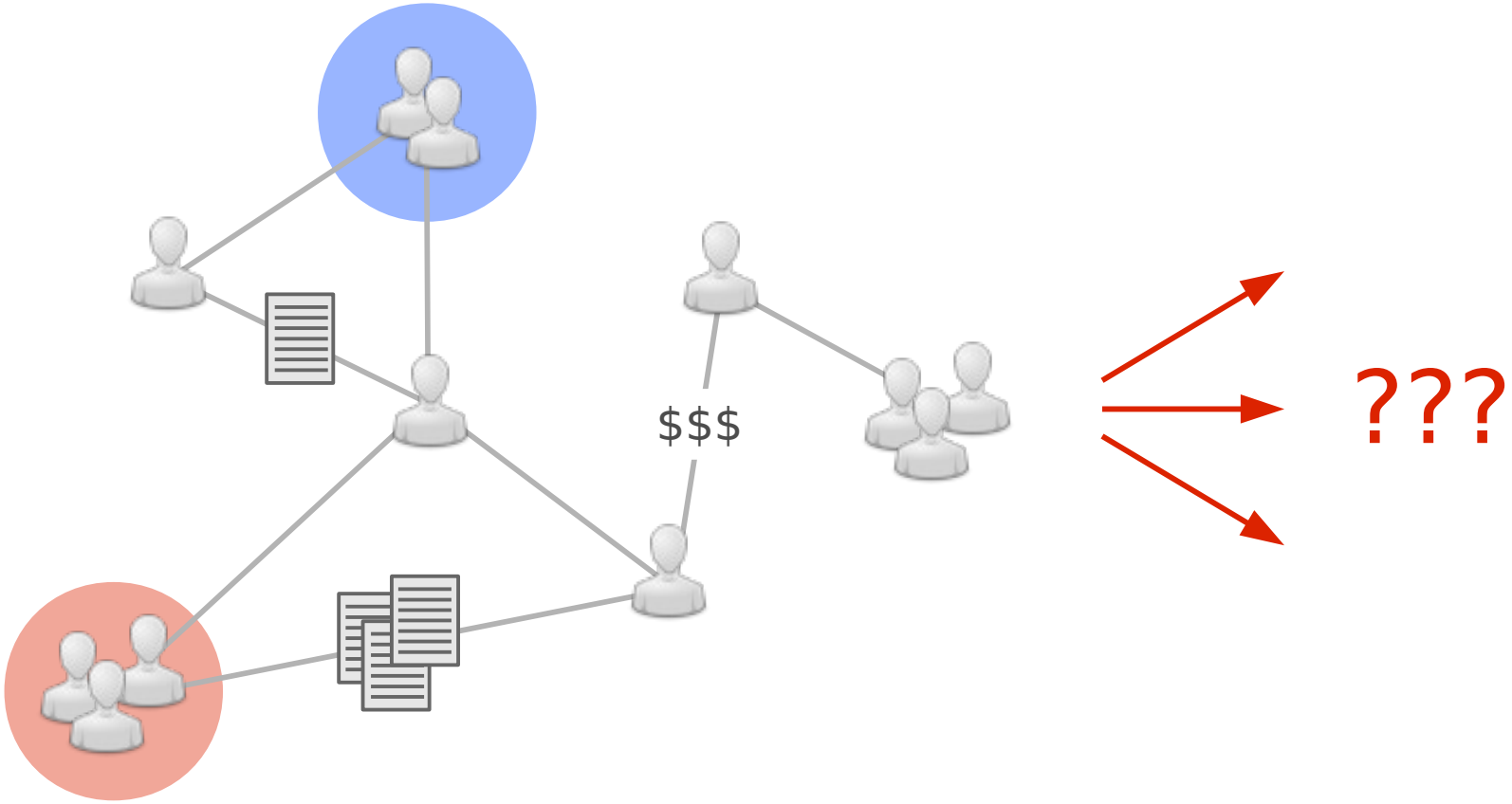
— NYT, 25 September 2007

Modeling Social Processes

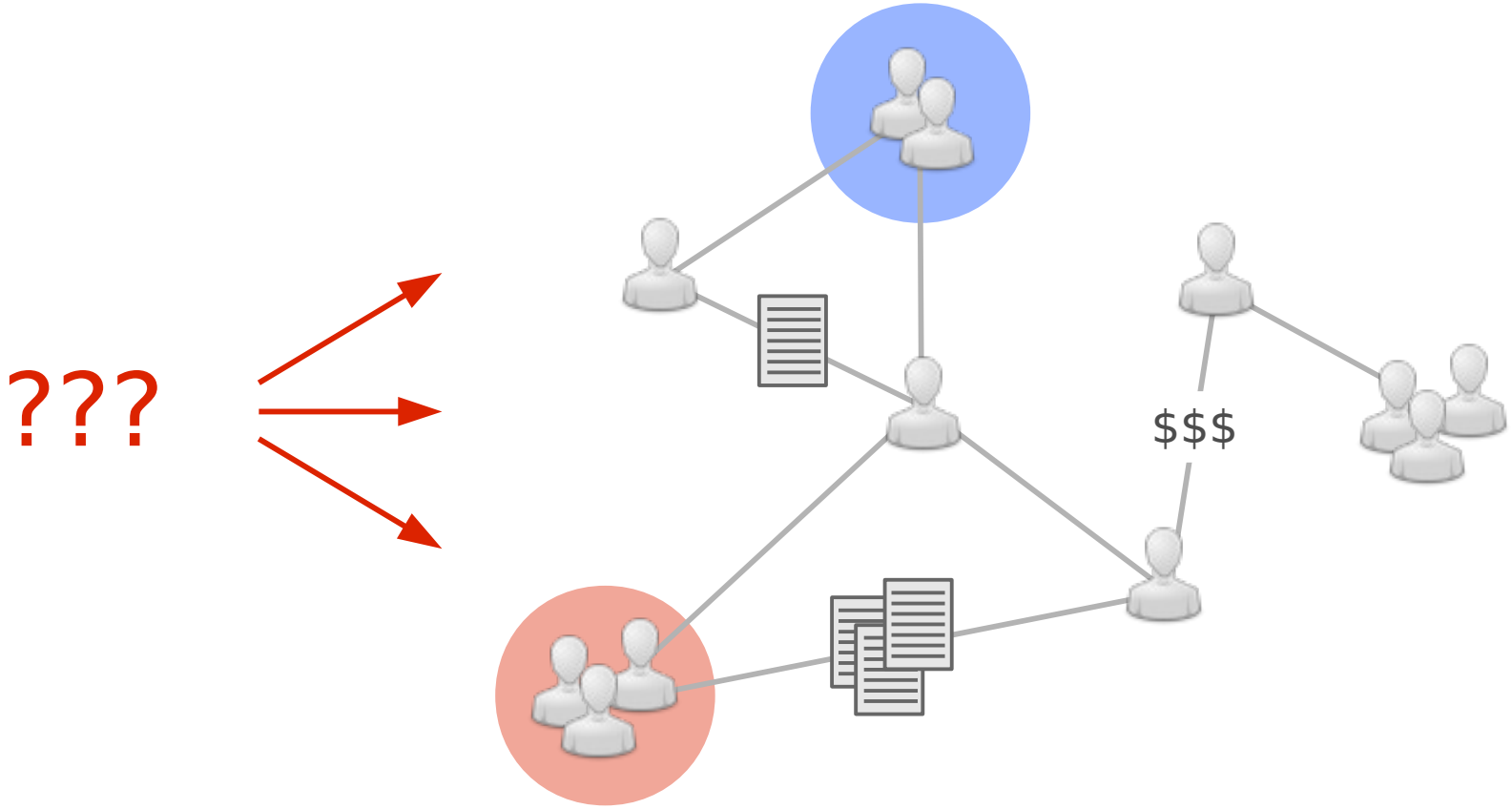
"Policy-makers or computer scientists may be interested in finding the needle in the haystack (such as a potential terrorist threat or the right web page to display from a search), but social scientists are more commonly interested in characterizing the haystack."

— King & Hopkins, 2010

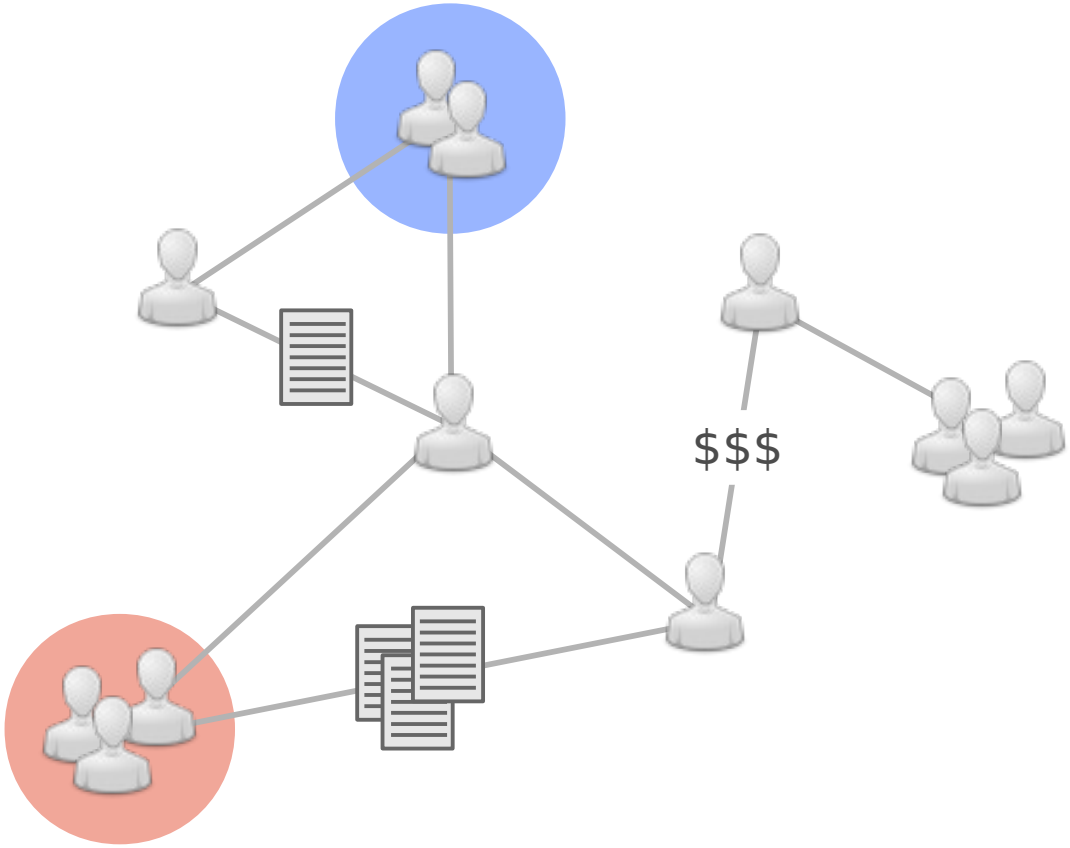
Predictive Analyses



Explanatory Analyses



Exploratory Analyses



= ???

Topics and Words

probability

gene	ncbi	computer	patent
genome	national	modeling	patenting
dna	information	data	claims
genetic	technology	algorithm	intellectual
genes	database	analyses	property
sequence	molecular	method	rights
human	biology	model	ip
protein	genbank	information	innovation
rna	pubmed	efficient	claim
genomic	references	complexity	claiming
...	...	...	...

...

Documents and Topics

POLICY FORUM

INTELLECTUAL PROPERTY

Intellectual Property Landscape of the Human Genome

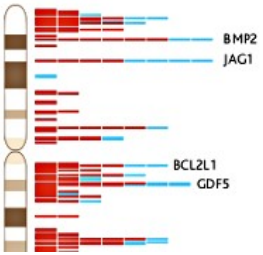
Kyle Jensen and Fiona Murray*

Gene patents are the subject of considerable debate and yet, like the term "gene" itself, the definition of what constitutes a gene patent is fuzzy (*1*). Nonetheless, gene patents that seem to cause the most controversy are those claiming human protein-encoding nucleotide sequences. This category is the subject of our analysis of the patent landscape of the human genome (*2*).

Enhanced online at www.sciencemag.org/cgi/content/full/310/5746/239

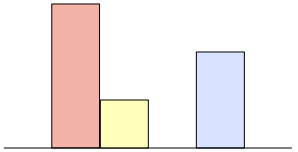
Critics describe the growth in gene sequence patents as an intellectual property (IP) "land grab" over a finite number of human genes (*3*, *4*). They suggest that overly broad patents might block follow-on research (*5*). Alternatively, gene IP rights may become highly fragmented and cause an anticommons effect, imposing high costs on future innovators and underuse of genomic resources (*6*). Both situations, critics argue, would increase the costs of genetic diagnostics, slow the development of new medicines, stifle academic research, tinguishing patents on the human genome from those on other species (*23*).

Our detailed map was developed using bioinformatics methods to compare nucleotide sequences claimed in U.S. patents to the human genome. Specifically, this map is based on a BLAST (*24*) homology search linking nucleotide sequences disclosed and claimed in granted U.S. utility patents to the set of protein-encoding messenger RNA transcripts contained in the National Center for Biotechnology Information (NCBI) RefSeq (*25*) and Gene (*26*) databases. This method allows us to map gene-oriented IP rights to specific physical loci on the human genome (*27*) (see figure, right). Our approach is highly specific in its identification of patents that actually claim human nucleotide sequences. However, by limiting the search to patents using the canoni-



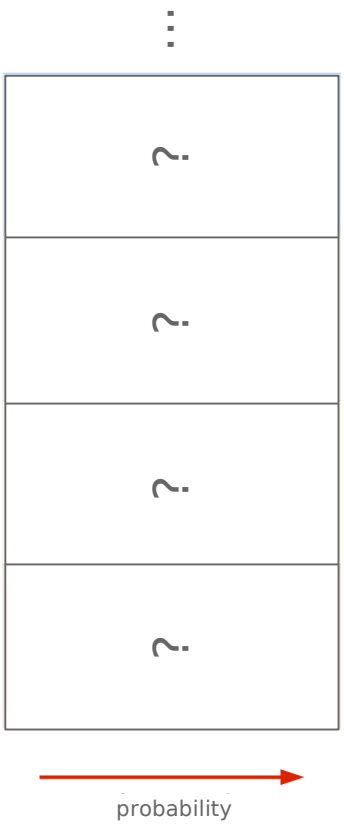
California, Isis Pharmaceuticals, the former SmithKline Beecham, and Human Genome Sciences. The top patent assignee is Incyte Pharmaceuticals/Incyte Genomics, whose IP rights cover 2000 human genes, mainly for use as probes on DNA microarrays.

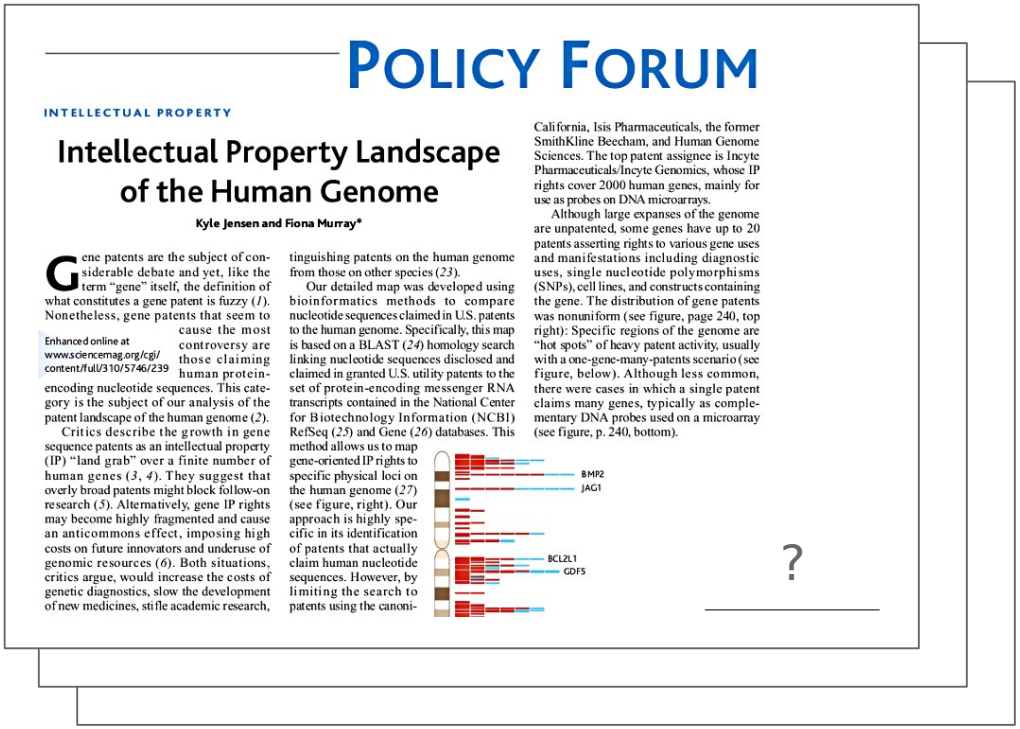
Although large expanses of the genome are unpatented, some genes have up to 20 patents asserting rights to various gene uses and manifestations including diagnostic uses, single nucleotide polymorphisms (SNPs), cell lines, and constructs containing the gene. The distribution of gene patents was nonuniform (see figure, page 240, top right): Specific regions of the genome are "hot spots" of heavy patent activity, usually with a one-gene-many-patents scenario (see figure, below). Although less common, there were cases in which a single patent claims many genes, typically as complementary DNA probes used on a microarray (see figure, p. 240, bottom).



Latent Dirichlet Allocation

[Blei, Ng & Jordan, '03]





POLICY FORUM

INTELLECTUAL PROPERTY

Intellectual Property Landscape of the Human Genome

Kyle Jensen and Fiona Murray*

Gene patents are the subject of considerable debate and yet, like the term "gene" itself, the definition of what constitutes a gene patent is fuzzy (*1*). Nonetheless, gene patents that seem to cause the most controversy are those claiming human protein-encoding nucleotide sequences. This category is the subject of our analysis of the patent landscape of the human genome (*2*).

Enhanced online at www.sciencemag.org/cgi/content/full/310/5746/239

Critics describe the growth in gene sequence patents as an intellectual property (IP) "land grab" over a finite number of human genes (*3*, *4*). They suggest that overly broad patents might block follow-on research (*5*). Alternatively, gene IP rights may become highly fragmented and cause an anticommons effect, imposing high costs on future innovators and underuse of genomic resources (*6*). Both situations, critics argue, would increase the costs of genetic diagnostics, slow the development of new medicines, stifle academic research, tinguishing patents on the human genome from those on other species (*23*).

Our detailed map was developed using bioinformatics methods to compare nucleotide sequences claimed in U.S. patents to the human genome. Specifically, this map is based on a BLAST (*24*) homology search linking nucleotide sequences disclosed and claimed in granted U.S. utility patents to the set of protein-encoding messenger RNA transcripts contained in the National Center for Biotechnology Information (NCBI) RefSeq (*25*) and Gene (*26*) databases. This method allows us to map gene-oriented IP rights to specific physical loci on the human genome (*27*) (see figure, right). Our approach is highly specific in its identification of patents that actually claim human nucleotide sequences. However, by limiting the search to patents using the canoni-

California, Isis Pharmaceuticals, the former SmithKline Beecham, and Human Genome Sciences. The top patent assignee is Incyte Pharmaceuticals/Incyte Genomics, whose IP rights cover 2000 human genes, mainly for use as probes on DNA microarrays.

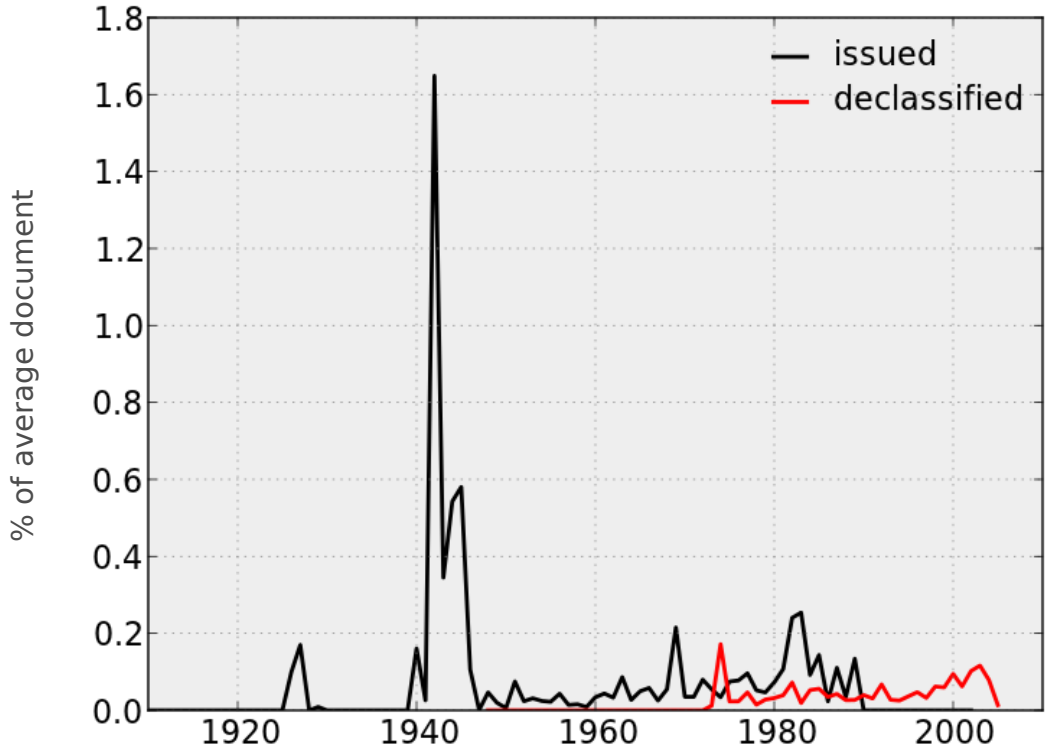
Although large expanses of the genome are unpatented, some genes have up to 20 patents asserting rights to various gene uses and manifestations including diagnostic uses, single nucleotide polymorphisms (SNPs), cell lines, and constructs containing the gene. The distribution of gene patents was nonuniform (see figure, page 240, top right): Specific regions of the genome are "hot spots" of heavy patent activity, usually with a one-gene-many-patents scenario (see figure, below). Although less common, there were cases in which a single patent claims many genes, typically as complementary DNA probes used on a microarray (see figure, p. 240, bottom).

BMP2
JAG1
BCL2L1
GDF5

?

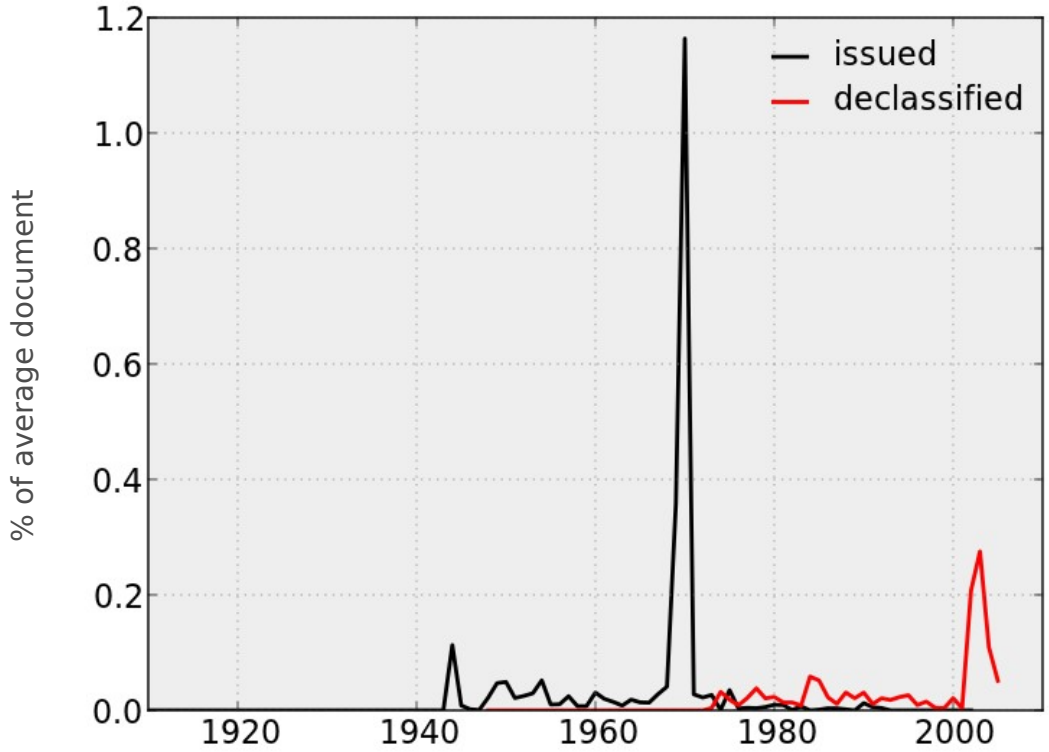
Inferred Topics

church
catholic
pope
vatican
religious
bishop
cardinal
archbishop
priests
paul
...



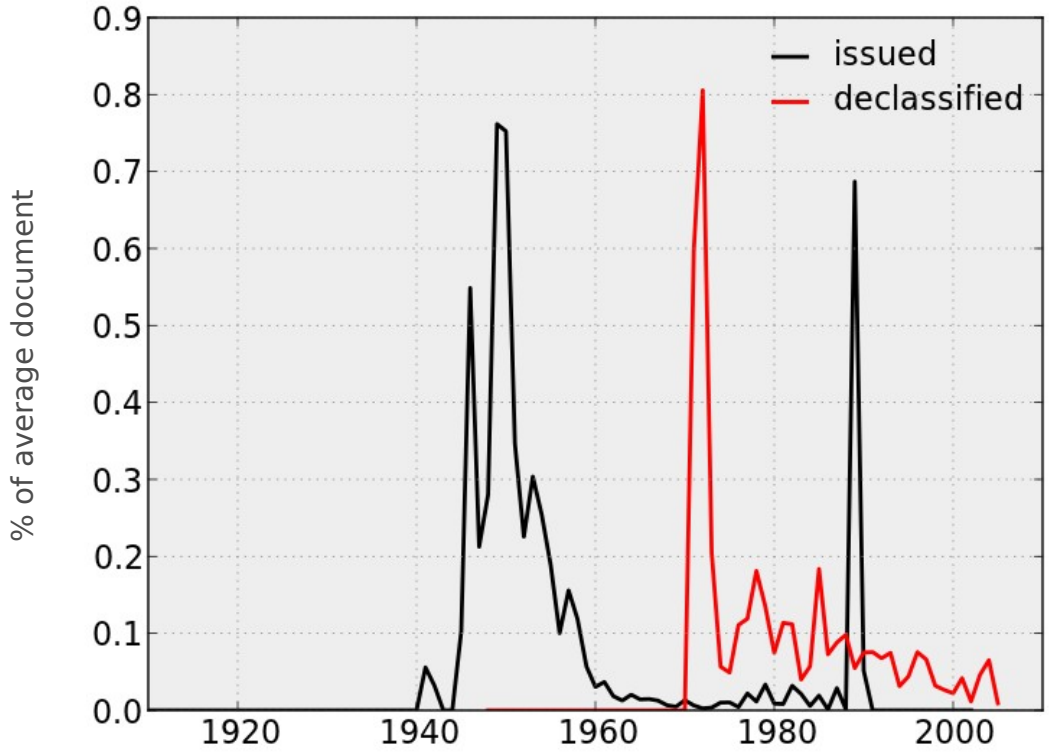
Inferred Topics

draft
service
manpower
volunteer
selective
age
calls
volunteers
deferments
pay
...

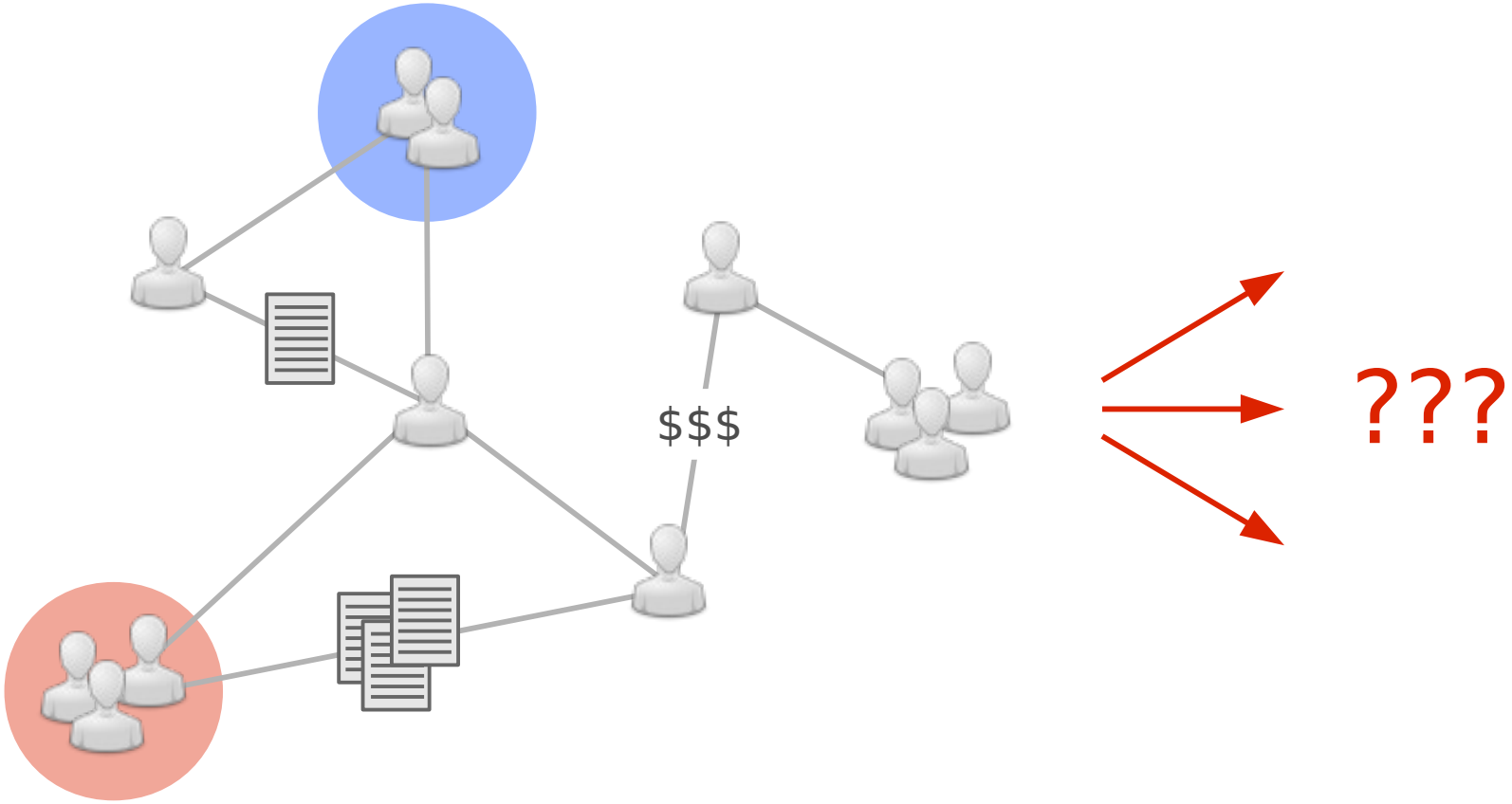


Inferred Topics

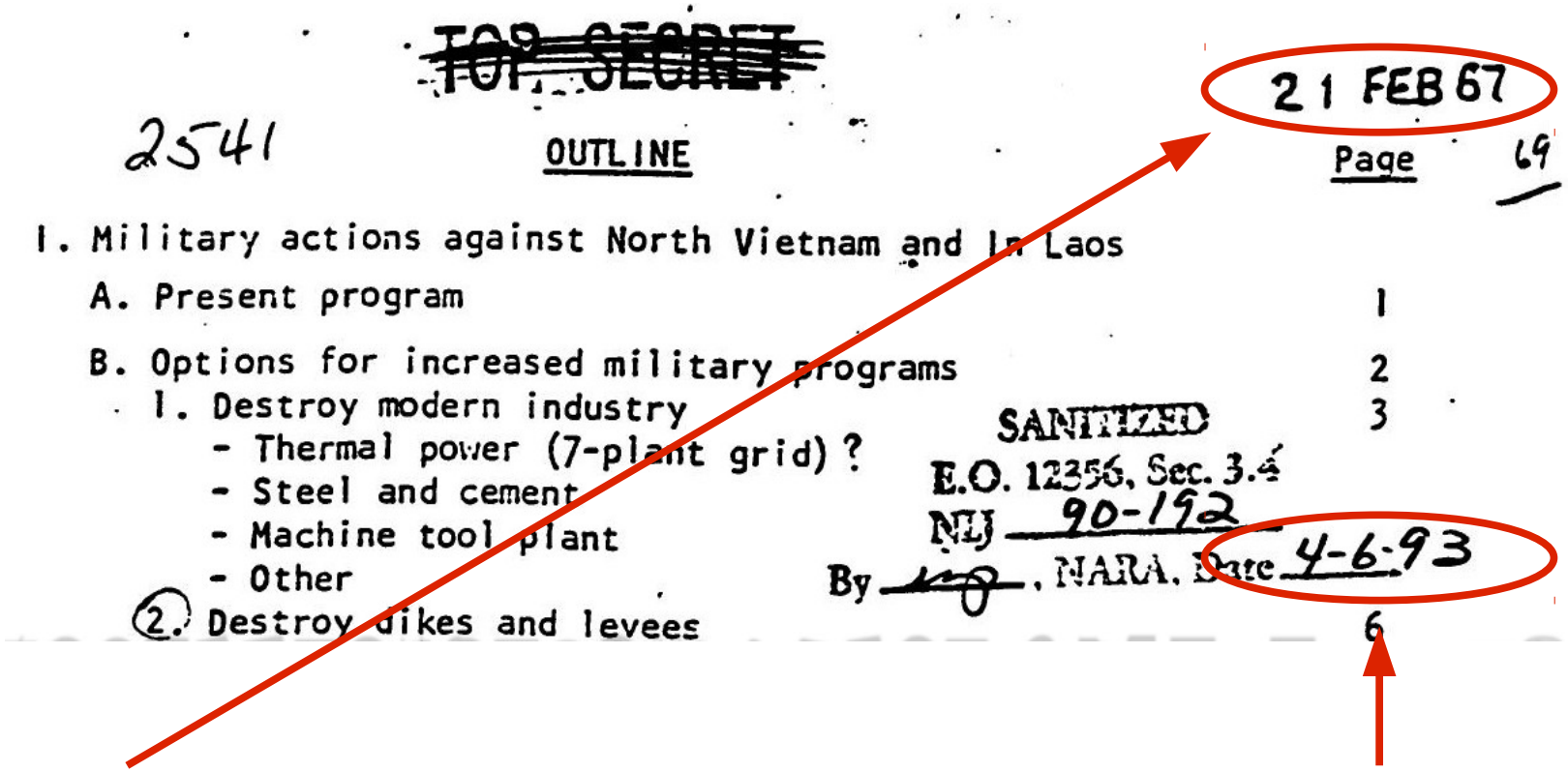
atomic
weapon
bomb
bombs
weapon
energy
thermonuclear
development
hydrogen
stockpile
...



Predictive Analyses



Classification Duration



~~TOP SECRET~~

2541

OUTLINE

21 FEB 67

Page 69

I. Military actions against North Vietnam and in Laos

A. Present program 1

B. Options for increased military programs 2

1. Destroy modern industry 3

- Thermal power (7-plant grid)?
- Steel and cement
- Machine tool plant
- Other

2. Destroy dikes and levees 6

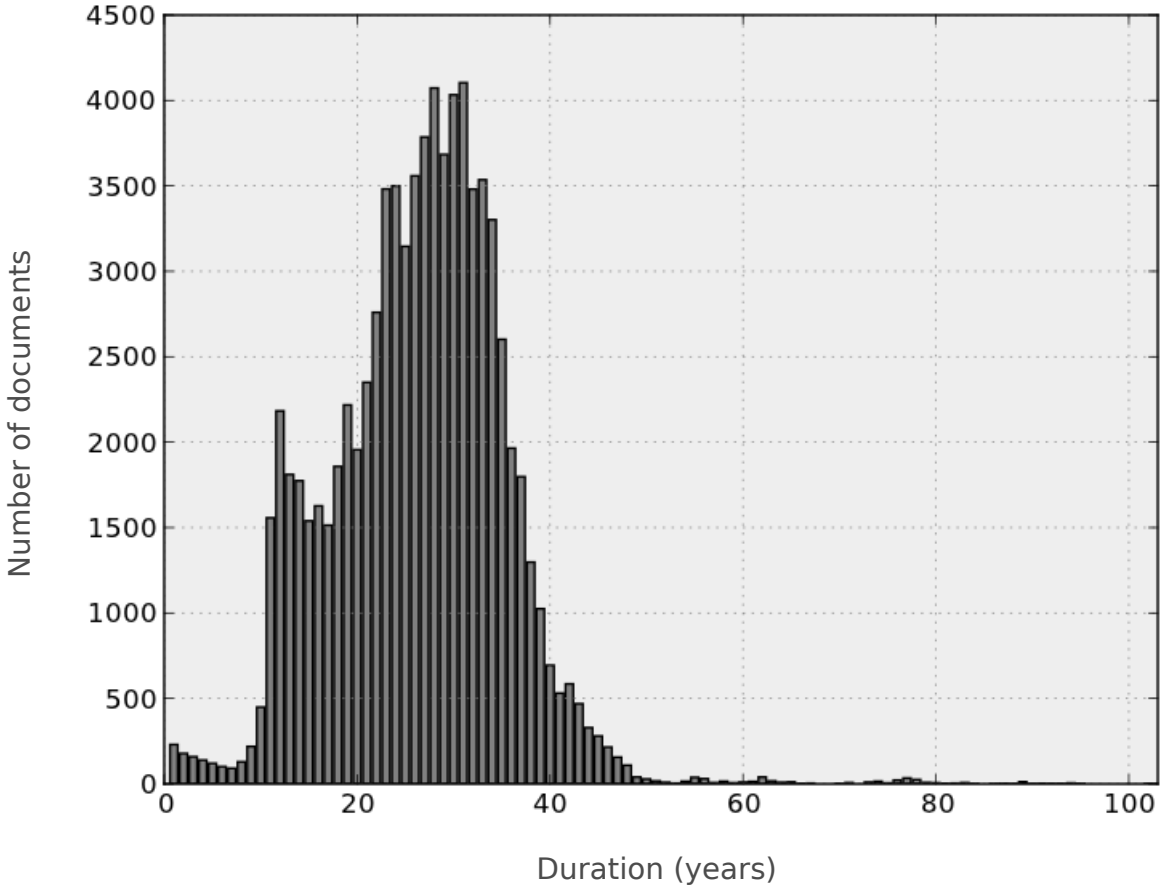
SANITIZED
E.O. 12356, Sec. 3.4
NLJ 90-192
By , NARA, Date 4-6-93

creation date
2/21/67

26 years

declassification date
4/6/93

Classification Durations



Survival Analysis

- Statistical methods for modeling durations:
  - Biology/medicine: organism death
  - Engineering: component failure
  - Social sciences: event durations (e.g., recidivism)
- Goal: model effect on survival time of covariates, e.g.,
  - Vaccine treatments
  - Temperature differences
  - Job placement or education programs

Duration and Content

HIS APPROACH WAS, "WELL, OF COURSE, WE KNOW THERE ISN'T ANYTHING TO THIS ALLEGED PHENOMENON (FLYING SAUCERS), BUT ON THE OTHER HAND". DURING HIS TALK SHKLOVSKIY AND OTHER SOVIETS JOKED AND LAUGHED AND OBVIOUSLY DID NOT TAKE THE SPEAKER'S REMARKS SERIOUSLY.

14 years

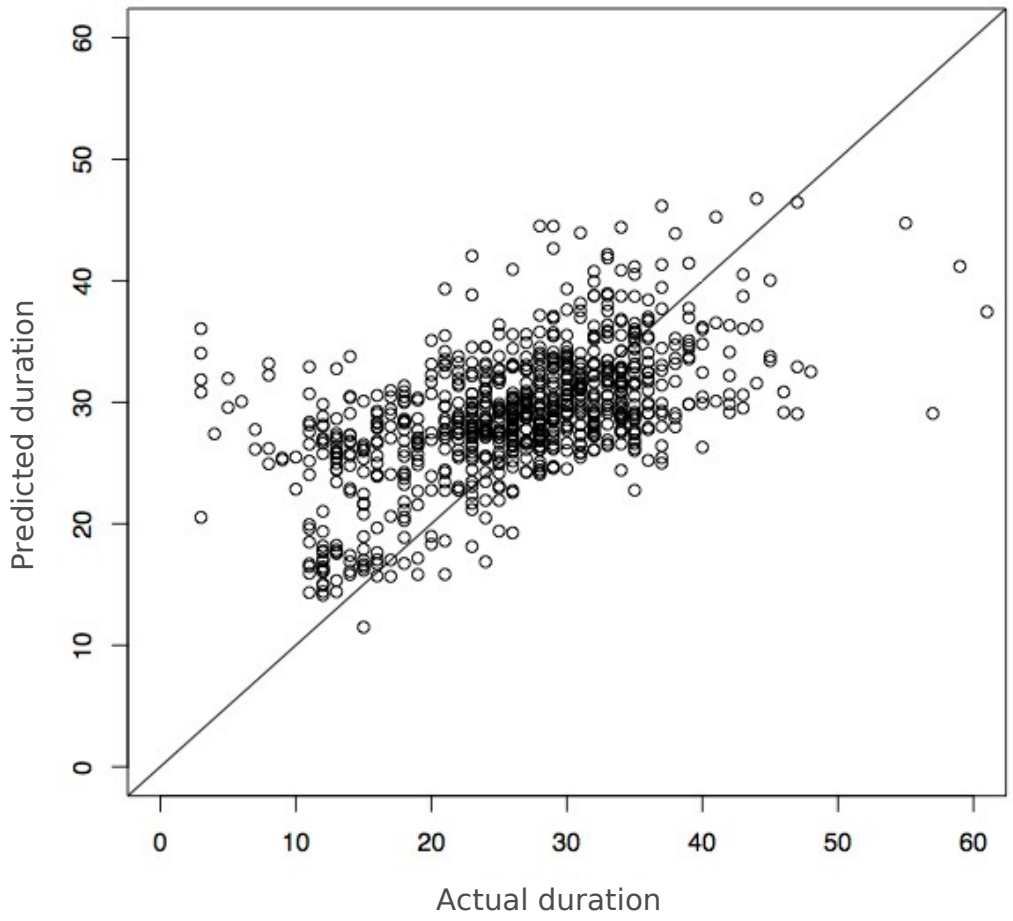
57 years

CENTRAL INTELLIGENCE GROUP

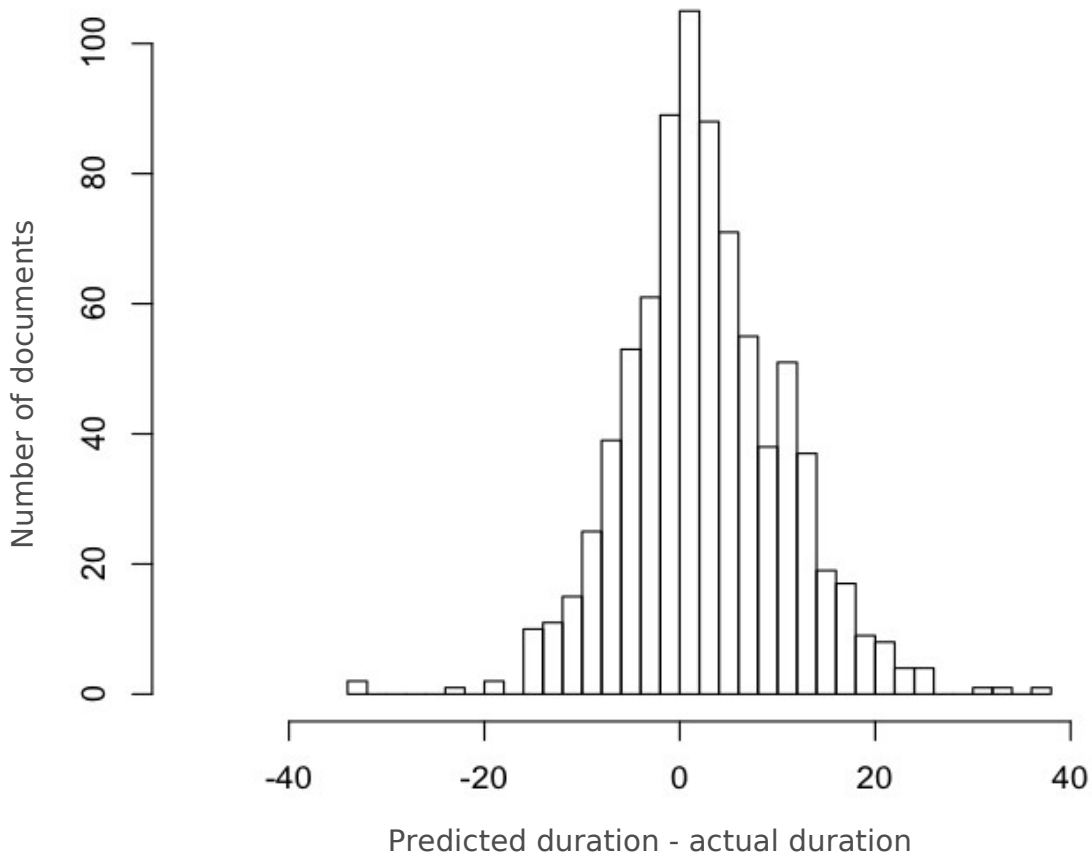
SOVIET CAPABILITIES FOR THE DEVELOPMENT AND PRODUCTION OF CERTAIN TYPES OF WEAPONS AND EQUIPMENT

1. Herein is presented an estimate of Soviet capabilities in the development and production, during the next ten years, of certain weapons and equipment, as follows:

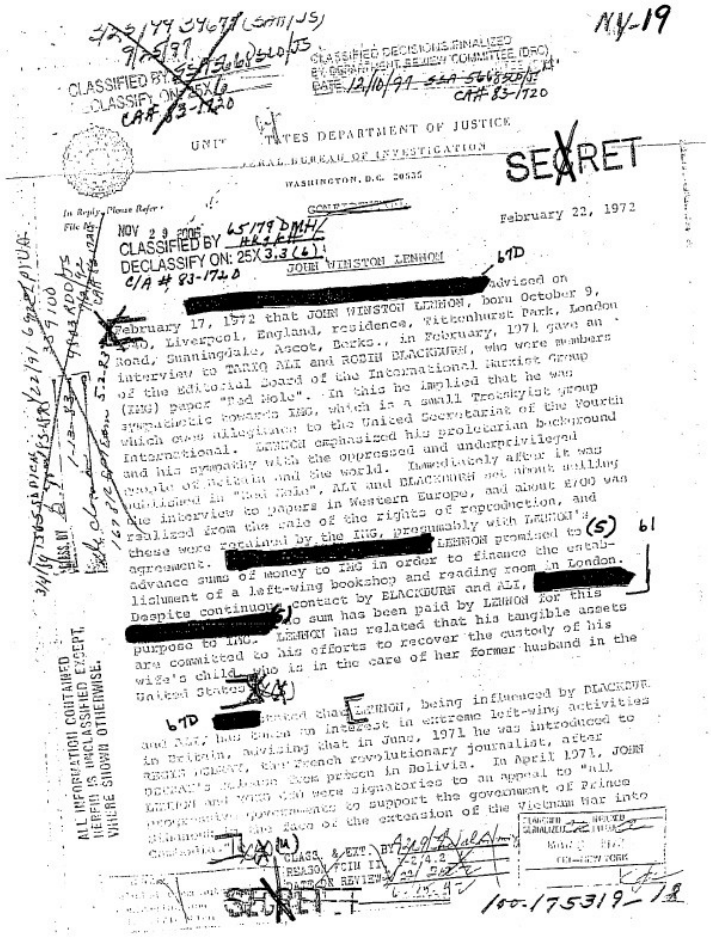
Predicting Duration Using Topics



Error Analysis



Modeling Text and Duration


NY-19

CLASSIFIED DECISIONS FINALIZED BY DECLASSIFICATION REVIEW COMMITTEE (DRC) DATE 12/10/94

CLASSIFIED BY
DECLASSIFY ON: EX 6
CA# 83-1720

UNITED STATES DEPARTMENT OF JUSTICE
FEDERAL BUREAU OF INVESTIGATION
WASHINGTON, D.C. 20535

SECRET

In Reply, Please Refer to File No.

February 22, 1972

NOV 29 2005
CLASSIFIED BY 65179 DMH/
DECLASSIFY ON: 25X 3.3(b)(1)
C/A # 83-1720

JOHN WINSTON LENNON

b7D [redacted] advised on February 17, 1972 that JOHN WINSTON LENNON, born October 9, [illegible], Liverpool, England, residence, Tittenhurst Park, London Road, Sunningdale, Ascot, Berks., in February, 1971 gave an interview to TARIQ ALI and ROBIN BLACKBURN, who were members of the Editorial Board of the International Marxist Group (IMG) paper "Red Mole". In this he implied that he was sympathetic towards IMG, which is a small Trotskyist group which owes allegiance to the United Secretariat of the Fourth International. LENNON emphasized his proletarian background and his sympathy with the oppressed and underprivileged people of Britain and the world. Immediately after it was published in "Red Mole", ALI and BLACKBURN met about selling the interview to papers in Western Europe, and about £700 was realized from the sale of the rights of reproduction, and these were retained by the IMG, presumably with LENNON's agreement. [redacted] LENNON promised to (5) b1 advance sums of money to IMG in order to finance the establishment of a left-wing bookshop and reading room in London. Despite continuous [redacted] contact by BLACKBURN and ALI, [redacted] no sum has been paid by LENNON for this purpose to IMG. LENNON has related that his tangible assets are committed to his efforts to recover the custody of his wife's child, [redacted] who is in the care of her former husband in the United States. (4)

b7D [redacted] stated that LENNON, being influenced by BLACKBURN and ALI, has taken an interest in extreme left-wing activities in Britain, advising that in June, 1971 he was introduced to REGIS DEBRAY, the French revolutionary journalist, after DEBRAY's release from prison in Bolivia. In April 1971, JOHN LENNON and YOKO ONO were signatories to an appeal to "all progressive governments to support the government of Prince Sihanouk in the face of the extension of the Vietnam War into Cambodia." (4)

CLASS. & EXT. BY
REASON FCIM II
DATE OF REVIEW

SECRET

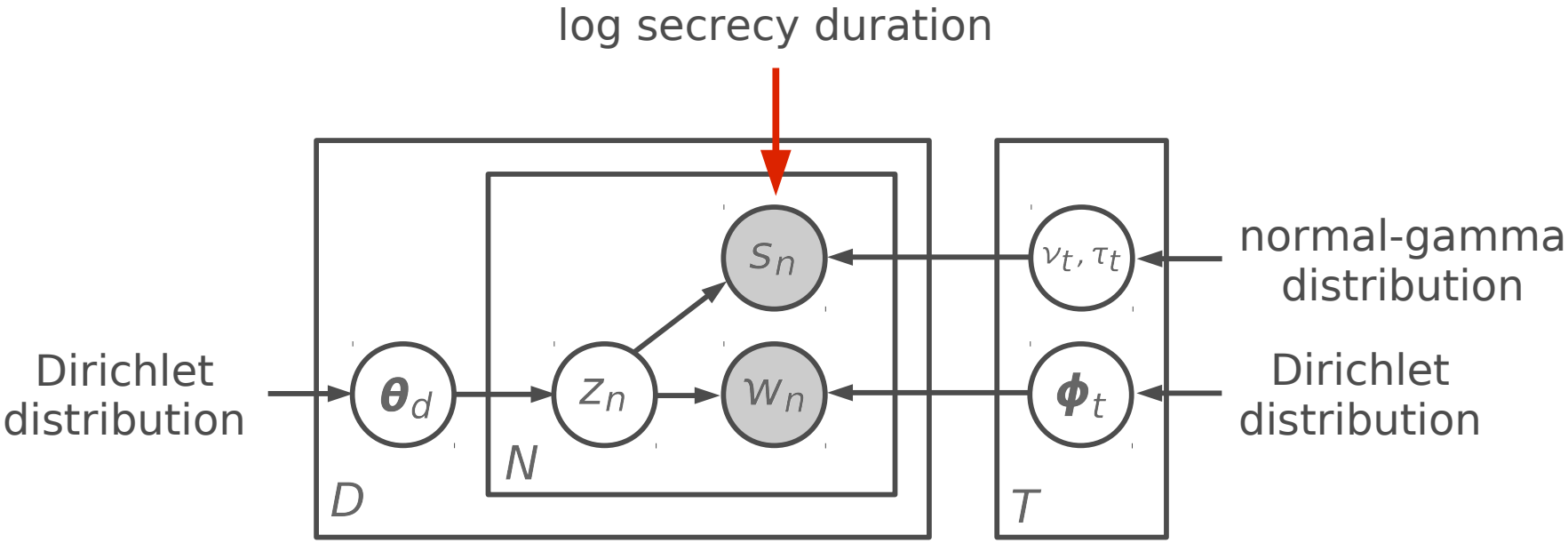
100-175319-18

ALL INFORMATION CONTAINED HEREIN IS UNCLASSIFIED EXCEPT WHERE SHOWN OTHERWISE.

- Topics provide information about classification durations
- Goal: incorporate durations into the probabilistic model
- Infer latent topics using both textual and temporal information

Jointly Modeling Text and Duration

[Shorey et al., '11]



Conclusions

- Government transparency is a complex social process
  - Structure, content, and dynamics
  - Exploratory, predictive, and explanatory analyses

- Exciting research area, many unexplored directions
- Progress requires interdisciplinary collaboration!

Thanks!

Acknowledgements: B. Desmarais, R. Shorey

wallach@cs.umass.edu
http://www.cs.umass.edu/~wallach/