Graphical Models for Text Analysis and its Applications

Arindam Banerjee
banerjee@cs.umn.edu

Dept of Computer Science & Engineering
University of Minnesota, Twin Cities

ASIAS Tools & Technology Symposium

July 27-28, 2009

Aviation Safety Reports (NASA)

Data and Goals

- Textual reports of problems/anomalies

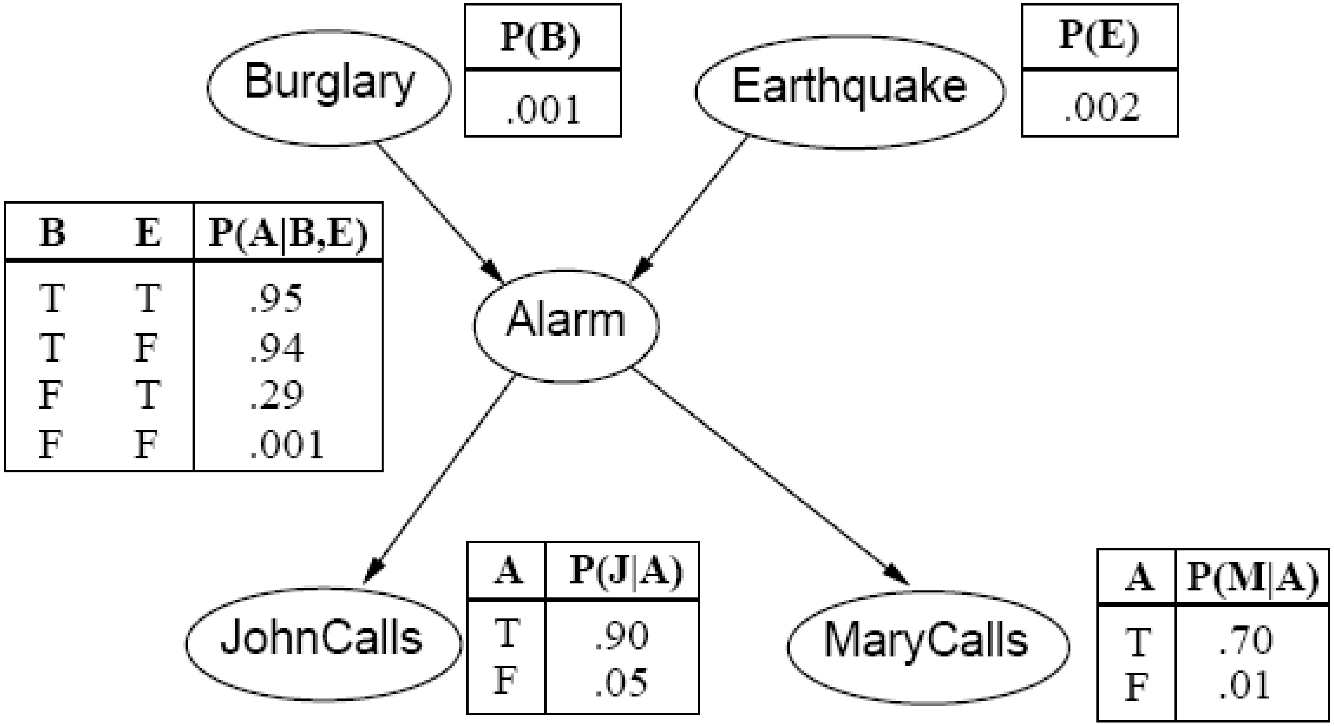
> I WAS FLYING THE KATANA WITH A STUDENT AND ON DOWNWIND **THE FUEL PRESSURE DROPPED TO ZERO, AND THE ENG WAS CUTTING OFF**. I VERIFIED FUEL PUMP WAS ON AND IT WAS ON. BY THE TIME WE TURNED SHORT FINAL, THE PROP STOPPED AND WE LANDED THE AIRPLANE SAFELY. THEN WE CALLED CASTLE UNICOM TO SEND THE FUEL TRUCK

- Topic Modeling:
  - Key topics discussed, types of events, etc.
  - Unsupervised analysis
- Text Classification:
  - Given a report, what is its anomaly/problem category
  - Supervised analysis
    - Use past category labeled reports to train

Graphical Models: What and Why

- Statistical Machine Learning
  - Build diagnostic/predictive models from data
  - Uncertainty quantification based on (minimal) assumptions
- The I.I.D. assumption
  - Data is independently and identically distributed
  - Example: Words in a doc are drawn i.i.d. from the dictionary
- Graphical models
  - Assume (graphical) dependencies between (random) variables
  - Closer to reality, domain knowledge can be captured
  - Learning/inference is much more difficult
- Bayesian Networks (BN)
  - *Directed* graphs, causal dependency

Example I: Burglary Network



P(B)
.001

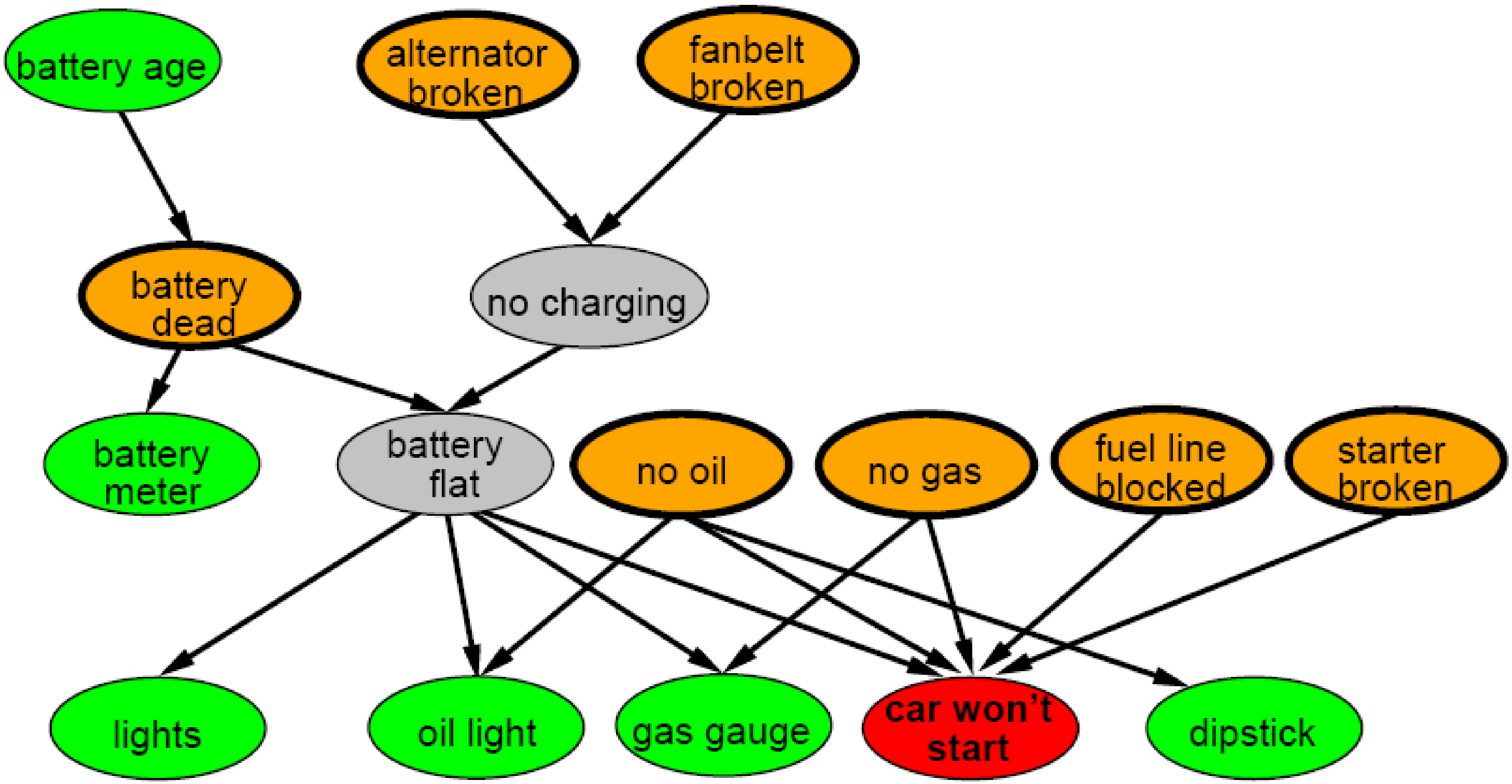
P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

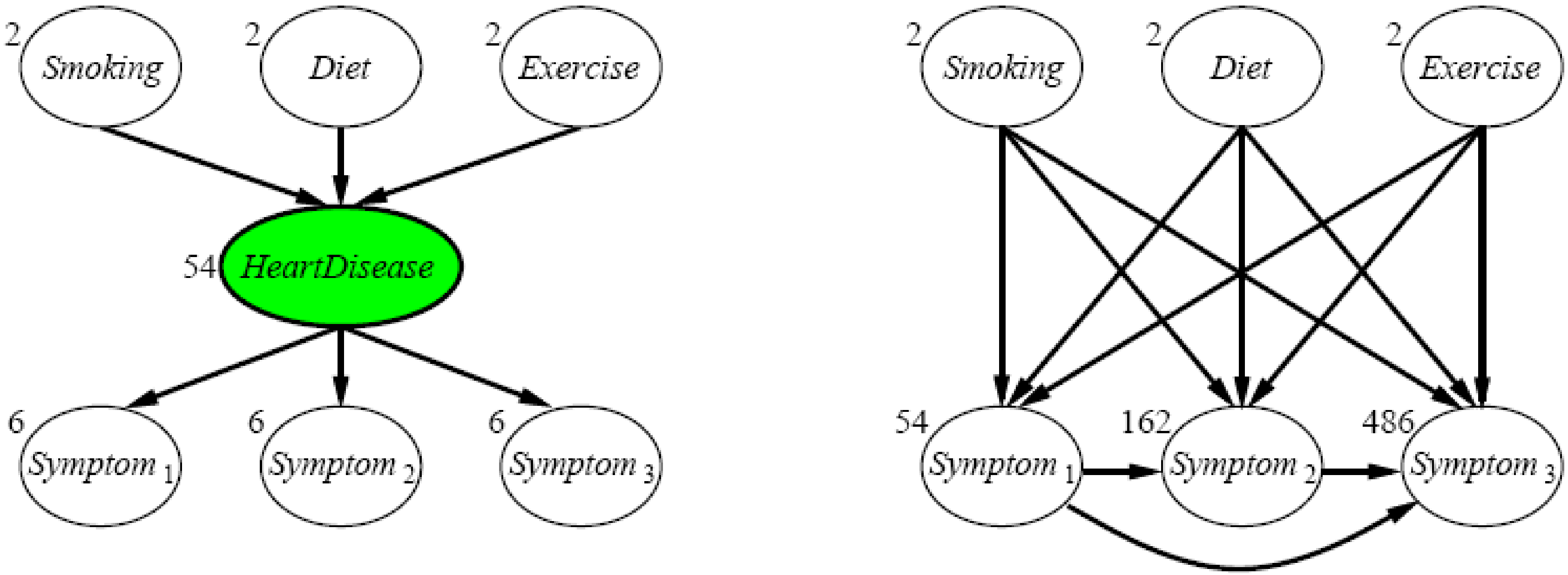
A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Example II: Car Problem Diagnosis



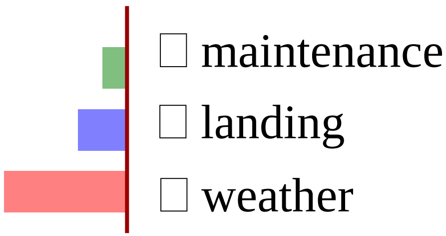
Latent Variable Models



- Bayesian network with hidden variables
  - Semantically more accurate, less parameters
- Example: Compute probability of heart disease

Topic Models

Document is a mixture of topics



Topic is a distribution over words

Maintenance: *check* *gear* *fuel* …
(0.02 0.01 0.01 …)

Landing: *undercarriage* *height* *runway* …
(0.025 0.02 0.01 …)

Weather: *fog* ice *snow* …
(0.04 0.03 0.02 …)

To generative a word: (i) Pick a topic, (ii) Sample a word

ice .. **frozen** .. **runway**

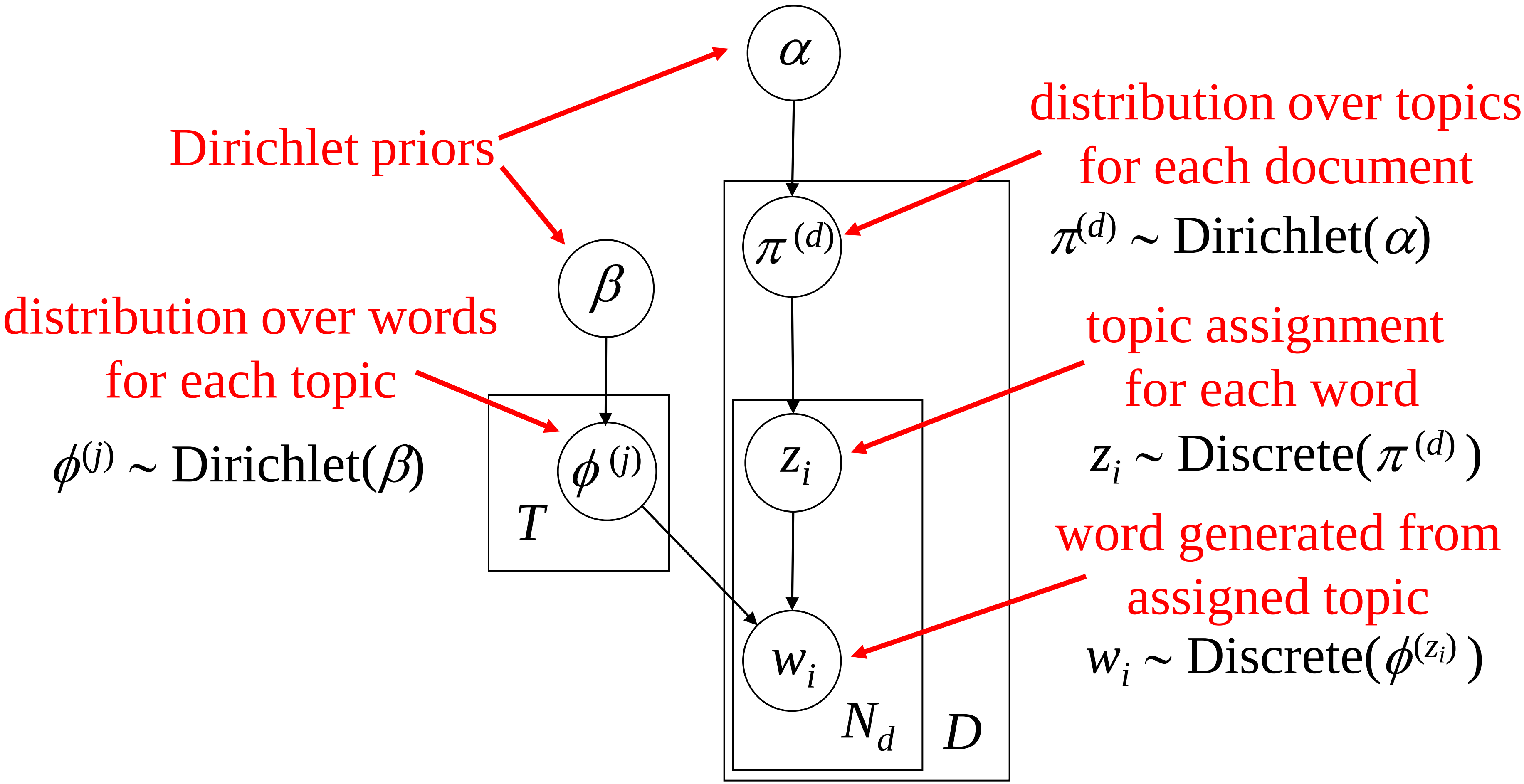
Example I: Topics in Slashdot

music	web	scientists	internet	games
apple	google	nasa	broadband	gaming
itunes	search	space	domain	game
riaa	yahoo	researchers	net	nintendo
ipod	site	science	network	sony
wikipedia	online	years	verisign	xbox
digital	sites	earth	bittorrent	gamers
napster	ebay	found	icann	wii
file	amazon	brain	service	console
drm	engine	university	access	video
songs	users	human	voip	article
industry	browser	research	dns	microsoft

Example II: Topics in Newsgroups

windows	turkish	game	god	israeli
dos	armenian	team	bible	israel
files	armenia	games	christian	moral
file	genocide	hockey	jesus	arabs
disk	turkey	year	church	arab
drive	radar	play	christians	absolute
port	armenians	season	atheism	killed
program	soviet	baseball	religion	morality
irq	list	pens	people	lebanon
ftp	turks	players	faith	lebanese
modem	detector	league	life	people
ibm	people	player	christianity	civilians

Latent Dirichlet Allocation (LDA)



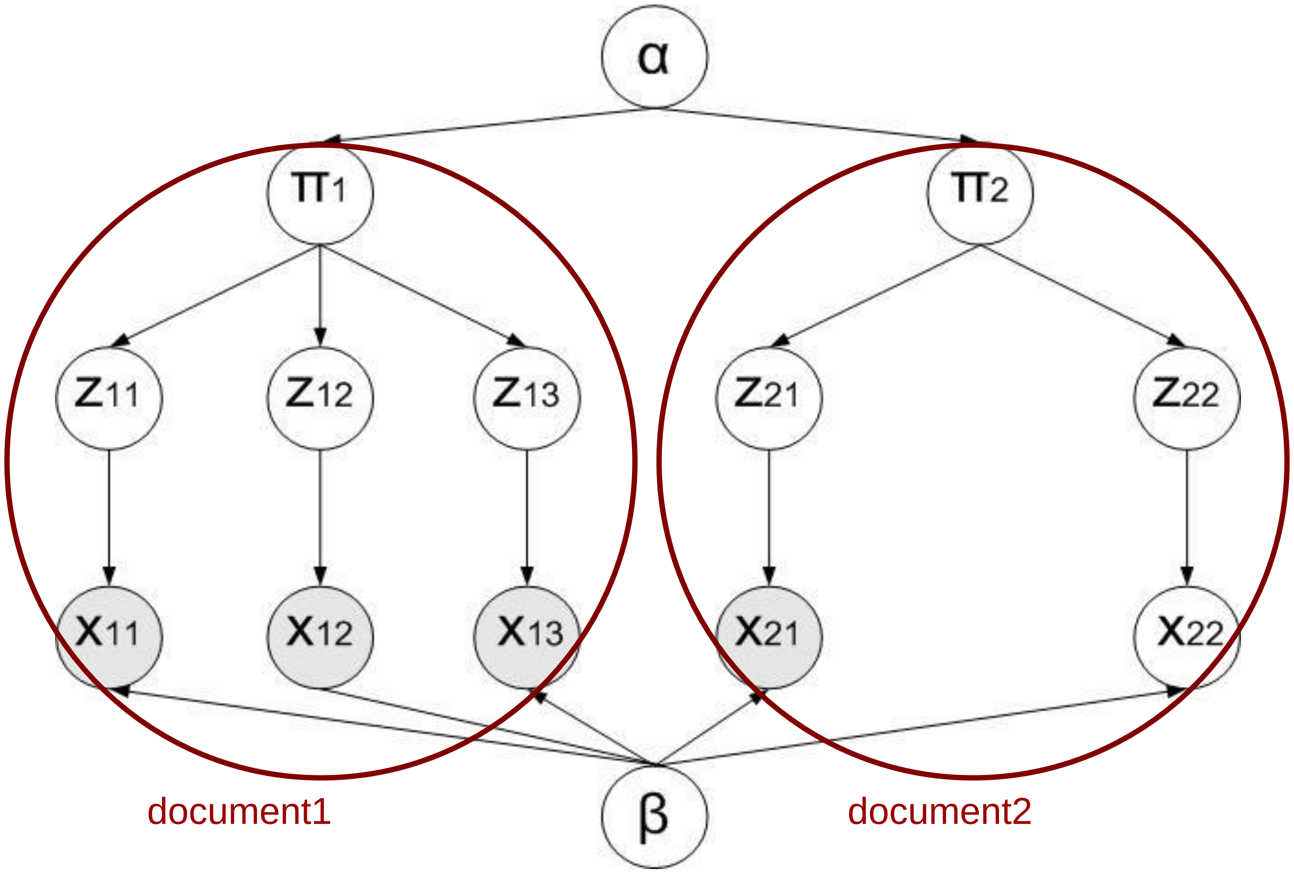
$\pi^{(d)} \sim \text{Dirichlet}(\alpha)$

$\phi^{(j)} \sim \text{Dirichlet}(\beta)$

$z_i \sim \text{Discrete}(\pi^{(d)})$

$w_i \sim \text{Discrete}(\phi^{(z_i)})$

LDA Generative Model: 2 Documents

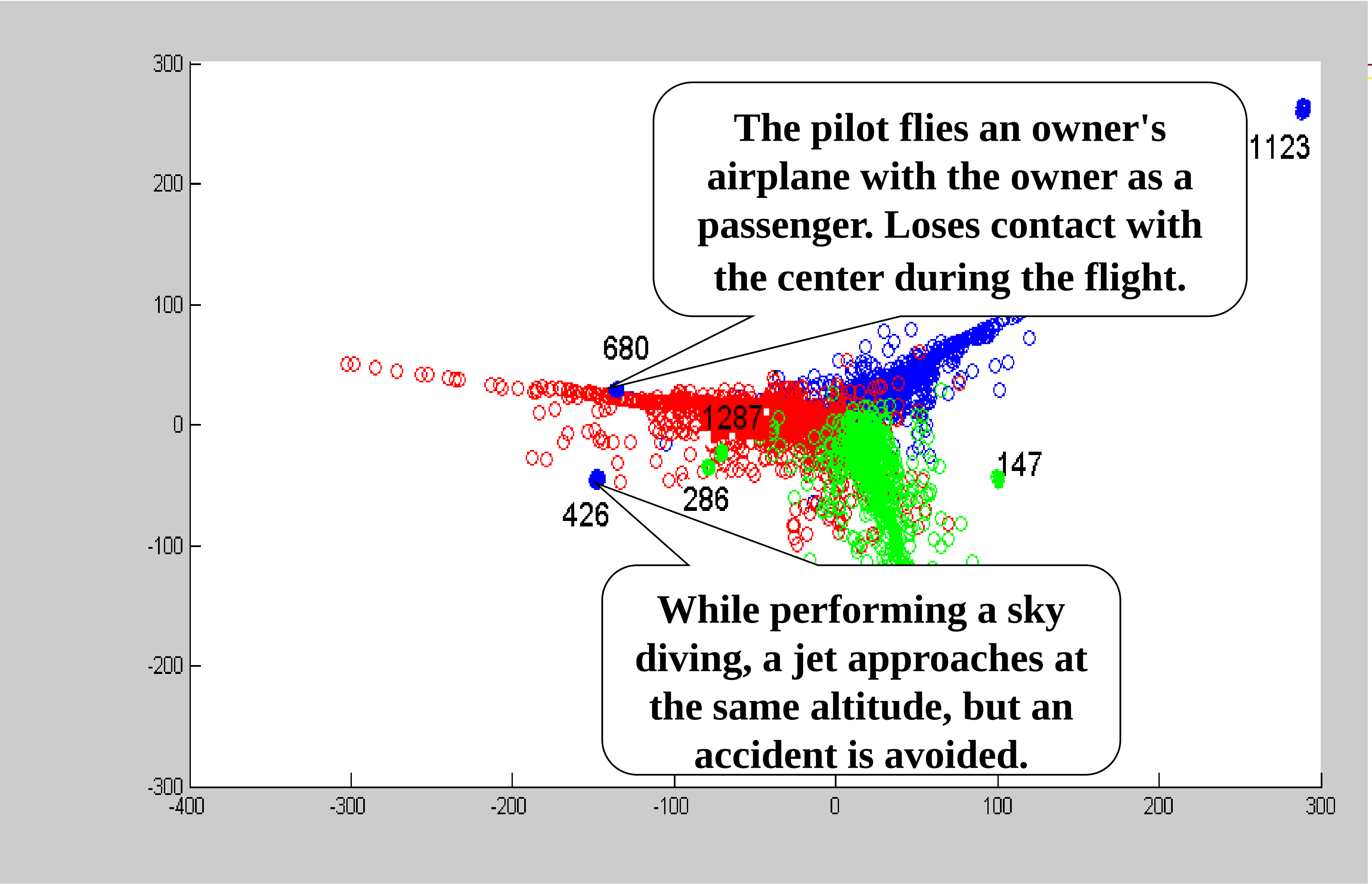


Results: Topics in ASRS Reports

A set of reports from ASRS: problems related to
(i) Flight crew performance, (ii) Passenger problems, (iii) Maintenance issues

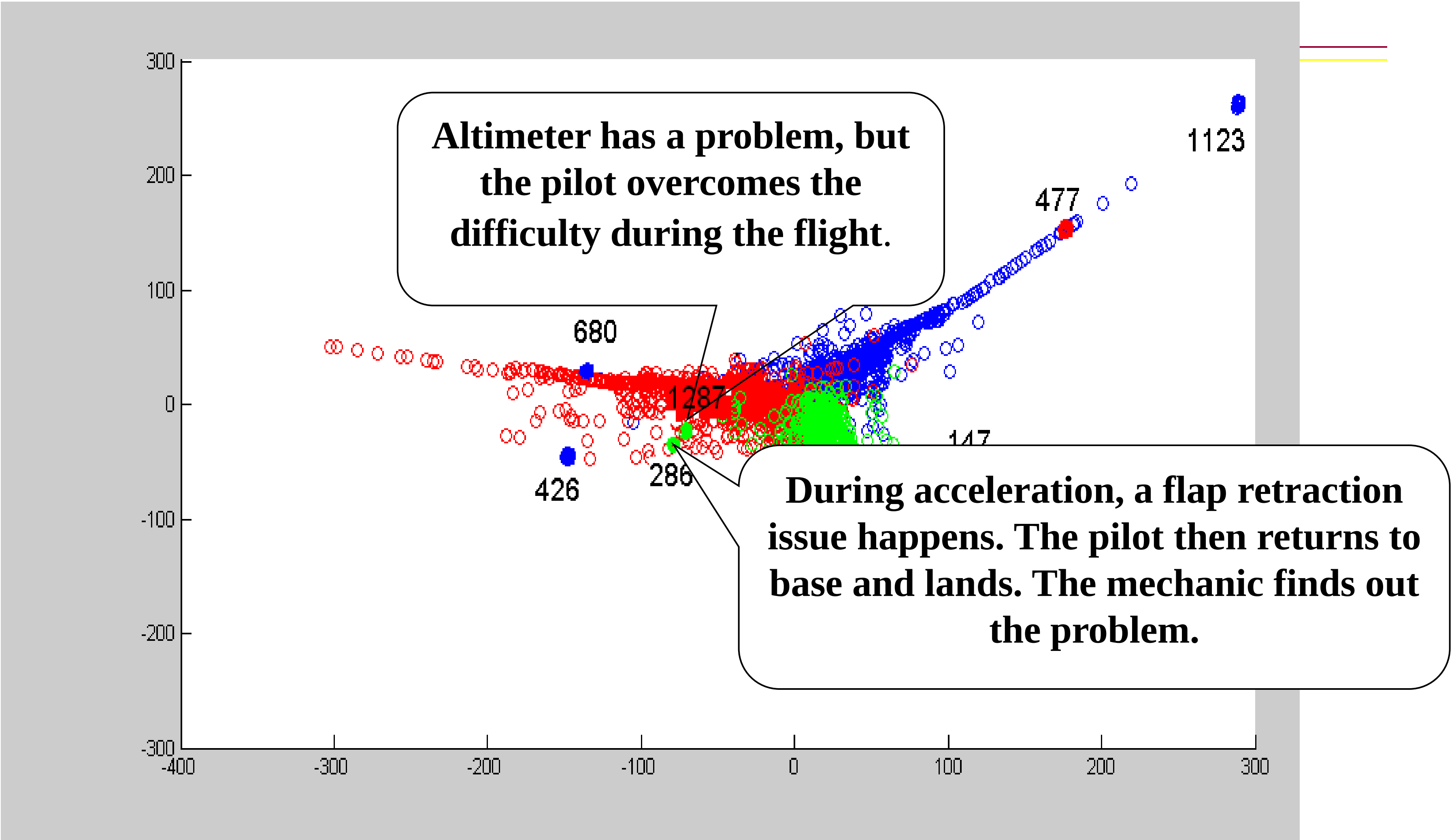
Flight Crew	Passenger	Maintenance
runway	passenger	aircraft
approach	flight	maintenance
aircraft	attendant	engine
departure	captain	zzz
altitude	seat	flight
turn	told	minimum equipment list
time	asked	check
atc	back	fuel
flight	attendants	time
tower	aircraft	gear

Two-Dimensional Visualization for Reports



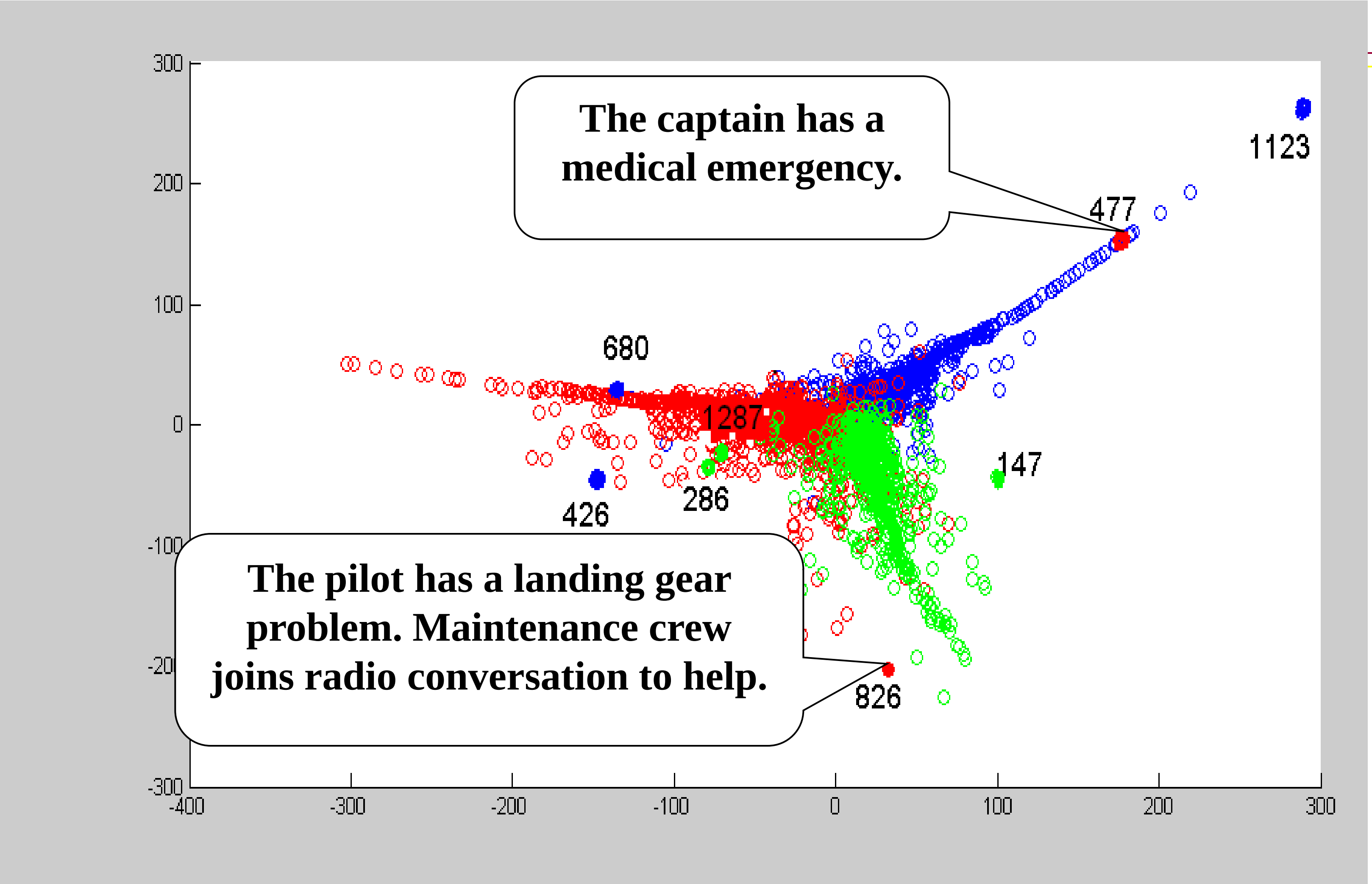
Red: Flight Crew Blue: Passenger Green: Maintenance

Two-Dimensional Visualization for Reports



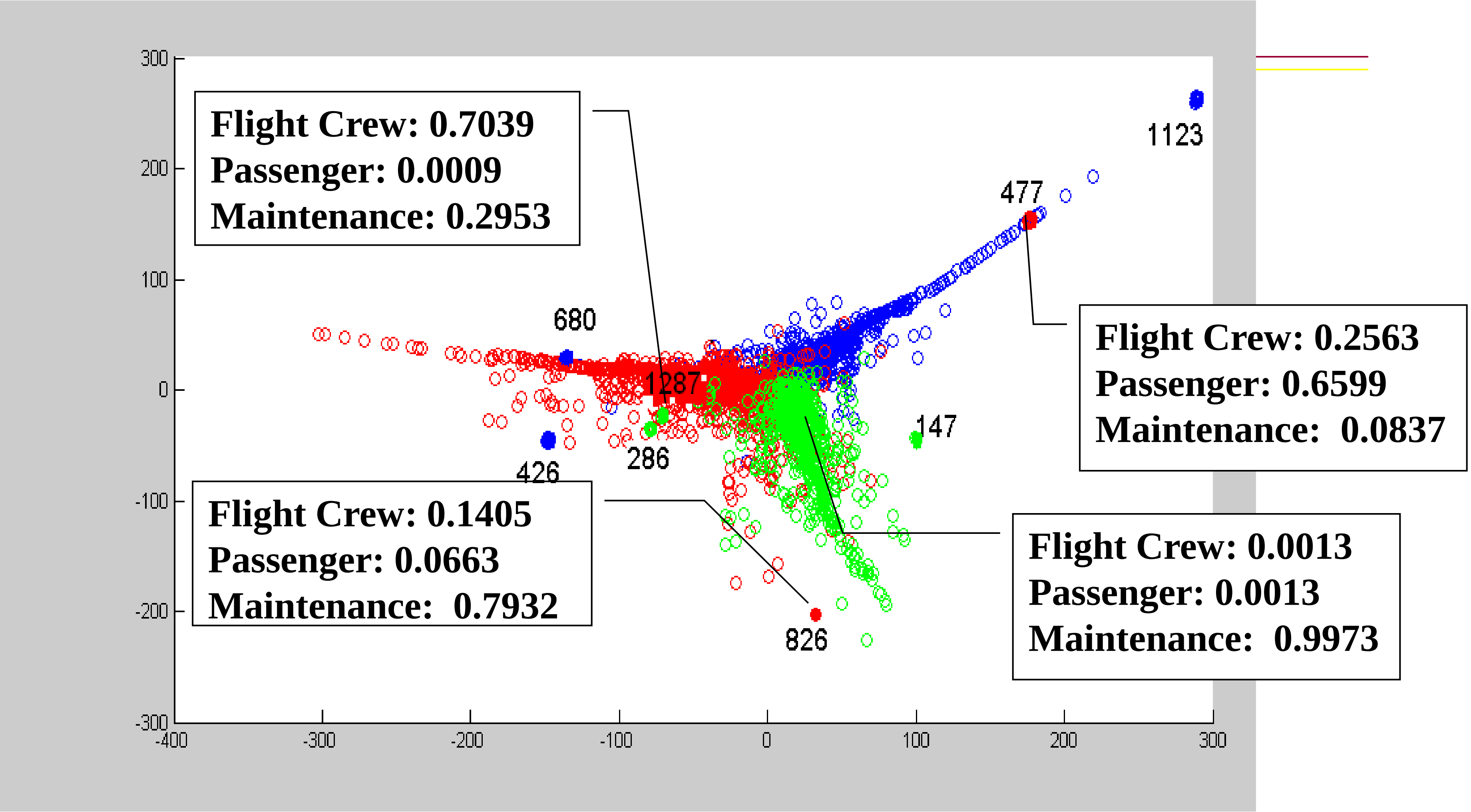
Red: Flight Crew　　Blue: Passenger　　Green: Maintenance

Two-Dimensional Visualization for Reports



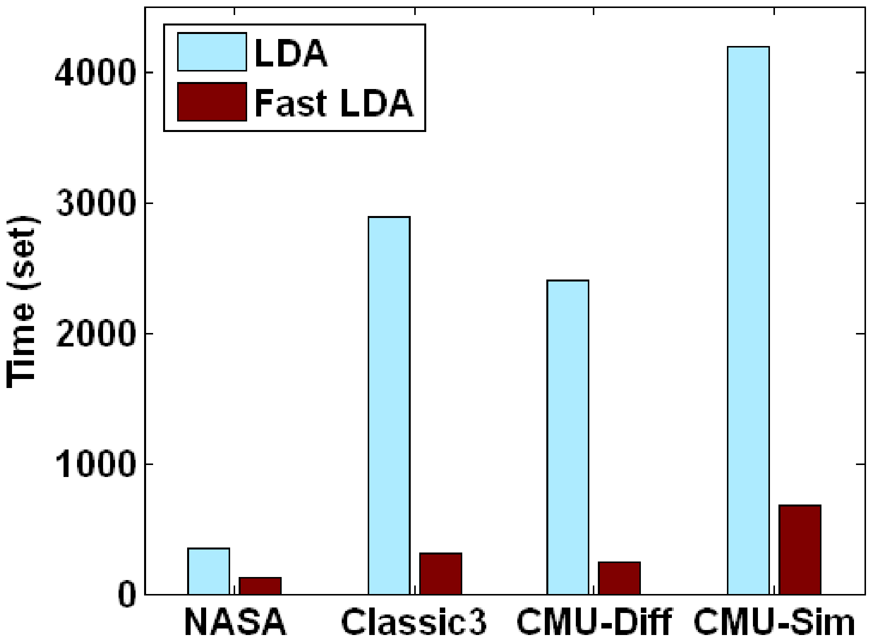
Red: Flight crew Blue: Passenger Green: Maintenance

Mixed Membership of Reports

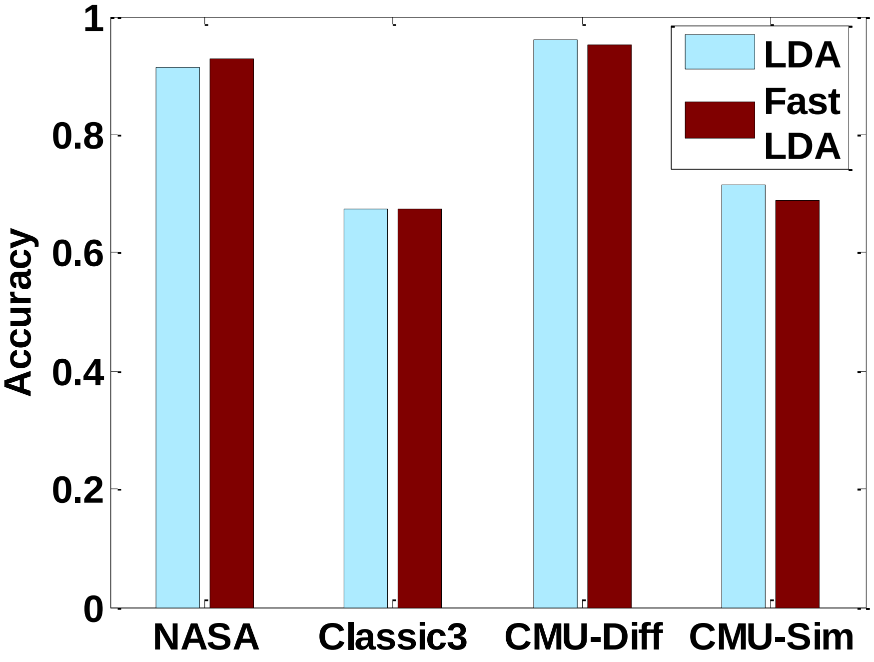


Red: Flight Crew Blue: Passenger Green: Maintenance

LDA vs FastLDA



Time comparison



Accuracy comparison

Text Classification

LDA & Fast LDA

mixed membership over topics

e.g. xxx did not work during landing because of the bad weather…

Maintenance (0.2) Landing (0.3) Weather(0.5)

interpretable topics

e.g. maintenance, mechanical, inspection…
Maintenance Problem

Discriminative Latent Dirichlet Allocation (DLDA)

SVM, logistic regression, etc.

Can use labels

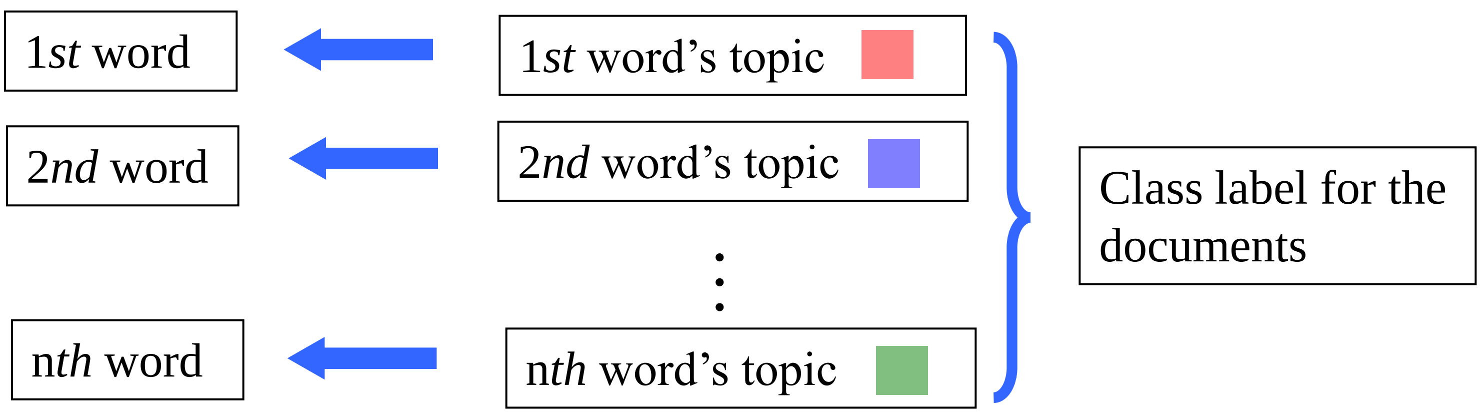
learn a classifier from documents & labels to classify new documents

no mixed membership

no interpretable topics

Discriminative LDA

- Supervised, learns classifier from training data
- Model generates documents and the labels
  - LDA for documents
  - Logistic Regression on topic proportions for labels
  - Number of topics independent of number of classes



Variational EM for DLDA: Overview

- Given: Documents (*X*), Labels (*Y*)
- Model: Parameters (Θ), Latent variables (*Z*)
- Maximum likelihood estimation of parameters

$$\Theta^* = \arg\max_{\Theta} \log p(X, Y \mid \Theta) = \arg\max_{\Theta} E[\log p(X, Y, Z \mid \Theta)]$$

- EM-based algorithm:
  - E-step: Use $p(Z|X, Y, \Theta)$ to compute $E[\log p(X,Y,Z|\Theta)]$
  - M-step: Compute Θ* which maximizes $E[\log p(X,Y,Z|\Theta)]$
- Issues
  - $p(Z|X, Y, \Theta)$ *cannot* be obtained in closed form
  - Computing $E[\log p(X,Y,Z|\Theta)]$ is *intractable*
- Variational Inference
  - Approximate $p(Z|X, Y)$ using $q(Z|\gamma)$
  - Choose γ to make $q(Z|\gamma) \sim p(Z|X,Y,\Theta)$

DLDA vs Others (NB,vMF,SVM,LR)

		Nasa	Classic3	Cmu-diff	Cmu-sim	Cmu-same
Fast DLDA with increasing # topics	Fast DLDA (c)	0.9237 ±0.0163	0.6756 ±0.0234	0.9800 ±0.0102	0.8653 ±0.0182	0.7900 ±0.0315
	Fast DLDA (c+15)	0.9232 ±0.0144	0.6858 ±0.0216	0.9747 ±0.0121	0.8713 ±0.0264	0.8458 ±0.0214
	Fast DLDA (c+30)	0.9301 ±0.0128	0.6838 ±0.0234	0.9817 ±0.0099	0.8707 ±0.0228	***0.8468 ±0.0190***
	Fast DLDA (c+50)	0.9237 ±0.0138	0.6854 ±0.0211	***0.9823 ±0.0083***	0.8700 ±0.0230	0.8150 ±0.0184
	Fast DLDA (c+100)	0.9261 ±0.0102	***0.6866 ±0.0245***	0.9760 ±0.0108	***0.8718 ±0.0182***	0.8347 ±0.0187
Generative models	vMF	0.9216 ±0.0113	0.6509 ±0.0246	0.9530 ±0.0071	0.7447 ±0.0214	0.7600 ±0.0347
	NB	***0.9334 ±0.0094***	0.6766 ±0.0230	0.9813 ±0.0069	0.8613 ±0.0216	0.8410 ±0.0262
Classification algorithms	LR	0.9209 ±0.0157	0.6396 ±0.0252	0.9553 ±0.0157	0.6750 ±0.1330	0.4823 ±0.1283
	SVM	0.9192 ±0.0146	0.6854 ±0.0278	0.9563 ±0.0105	0.8357 ±0.0156	0.8120 ±0.203

Larger # topics (k>c) usually => higher accuracy.

DLDA vs Others (NB,vMF)

		Nasa	Classic3	Cmu-diff	Cmu-sim	Cmu-same
Fast DLDA with increasing # topics	Fast DLDA (c)	0.9237 ±0.0163	0.6756 ±0.0234	0.9800 ±0.0102	0.8653 ±0.0182	0.7900 ±0.0315
	Fast DLDA (c+15)	0.9232 ±0.0144	0.6858 ±0.0216	0.9747 ±0.0121	0.8713 ±0.0264	0.8458 ±0.0214
	Fast DLDA (c+30)	0.9301 [illegible]	0.6838 [illegible]	0.9817 [illegible]	0.8707 [illegible]	*0.8468* [illegible]
	Fast DLDA (c+50)	[illegible] ±0.0138	[illegible] ±0.0211	[illegible] *±0.0083*	[illegible] ±0.0230	[illegible] ±0.0184
	Fast DLDA (c+100)	[illegible] ±0.0162	[illegible] *±0.0245*	[illegible] ±0.0108	[illegible] *±0.0182*	[illegible] ±0.0187
Generative models	vMF	0.9216 [illegible]	0.6509 [illegible]	0.9530 [illegible]	0.7447 [illegible]	0.7600 [illegible]
	NB	[illegible] *±0.0094*	[illegible] ±0.0230	[illegible] ±0.0069	[illegible] ±0.0216	[illegible] ±0.0262
Classification algorithms	LR	0.9209 ±0.0157	0.6396 ±0.0252	0.9553 ±0.0157	0.6750 ±0.1330	0.4823 ±0.1283
	SVM	0.9192 ±0.0146	0.6854 ±0.0278	0.9563 ±0.0105	0.8357 ±0.0156	0.8120 ±0.203

0.9301, 0.6866, 0.9823, 0.8718, 0.8468

∧ ∨ ∨ ∨ ∨

0.9334, 0.6766, 0.9813, 0.8613, 0.8410

p-value: 0.3328, 0.0161, 0.6709, 0.0365, 0.1128

DLDA vs Others (SVM,LR)

Fast DLDA with increasing # topics

Generative models

Classification algorithms

	Nasa	Classic3	Cmu-diff	Cmu-sim	Cmu-same
Fast DLDA (c)	0.9237 ±0.0163	0.6756 ±0.0234	0.9800 ±0.0102	0.8653 ±0.0182	0.7900 ±0.0315
Fast DLDA (c+15)	0.9232 ±0.0144	0.6858 ±0.0216	0.9747 ±0.0121	0.8713 ±0.0264	0.8458 ±0.0214
Fast DLDA (c+30)	0.9301	0.6838	0.9817	0.8707	*0.8468*
	0.9301,	**0.6866,**	**0.9823,**	**0.8718,**	**0.8468**
Fast DLDA (c+50)	±0.0138	±0.0211	±*0.0083*	±0.0230	±0.0184
Fast DLDA (c+100)	0.9261	*0.6866*	0.9760	*0.8718*	0.8347
	V	V	V	V	V
vMF	±0.0113	±0.0246	±0.0071	±0.0214	±0.0347
NB	*0.9334* ±*0.0094*	0.6766 ±0.0230	0.9813 ±0.0069	0.8613 ±0.0216	0.8410 ±0.0262
LR	0.9209	0.6396	0.9553	0.6750	0.4823
	0.9209,	**0.6854,**	**0.9563,**	**0.8357,**	**0.8120**
SVM	±0.0146	±0.0278	±0.0105	±0.0156	±0.203

p-value: 0.0087, 0.4205, 0.0025, <0.001, <0.001

DLDA vs LDA

	Nasa	Classic3	Cmu-diff	Cmu-sim	Cmu-same
Std LDA	0.9140 ±0.0140	0.6733 ±0.0254	0.9677 ±0.0069	0.8143 ± 0.0161	0.5633 ±0.0243
Fast LDA	0.9194 ±0.0148	0.6748 ± 0.0242	0.9773 ± 0.0110	0.8553 ±0.0197	0.7730 ±0.0205
Std DLDA	0.9220 ±0.0127	0.6710 ±0.0256	0.9600 ±0.0089	0.8140 ±0.0252	0.6267 ±0.0348
Fast DLDA	**0.9237 ±0.0163**	**0.6756 ±0.0234**	**0.9800 ±0.0102**	**0.8653 ±0.0182**	**0.7900 ±0.0315**

Topics inferred by DLDA

1	runway, aircraft, approach, tower, cleared, landing, airport, turn, taxi, traffic, final, controller	**Flight crew**
2	maintenance, aircraft, flight, minimum equipment list, time, check, engine, mechanical, installed, part, inspection, work	**Maintenance**
3	passenger, flight, attendant, told, captain, seat,asked, back, attendants, aircraft, lavatory, crew	**Passenger**
4	passenger, flight, medical, attendant, emergency, aircraft doctor, landing, attendants, captain, oxygen,paramedics	**Passenger Medical Emergency**

- First three topics correspond to three classes respectively
- Topic 4 is a subclass of class (3)

Summary

- LDA and FastLDA
  - Topic discovery from documents
  - Efficient algorithms, interpretable results, visualization
- Discriminative LDA
  - Text classification using topic models
  - Competitive with state-of-the-art (SVM,NB)
  - More interpretable
- Future Work
  - Leverage supplemental information in ASRS data
    - E.g., day/time, location, airport, time of year, equipment, etc.
  - Multi-category prediction
    - A document may report multiple different porblems

References

- A. Banerjee, H. Shan, *Latent Dirichlet Conditional Naïve Bayes Models*, ICDM, 2007.
- D. Blei, A. Ng, M. Jordan, *Latent Dirichlet Allocation*, JMLR, 2003.
- D. Blei, J. McAuliffe, *Supervised Topic Models*, NIPS, 2007.
- H. Shan, A. Banerjee, *Bayesian Coclustering*, ICDM, 2008.
- H. Shan, A. Banerjee, *Mixed Membership Naïve Bayes Models*, TR-09-002, Dept of CSE, University of Minnesota, Twin Cities.
- H. Shan, A. Banerjee, *Discriminative Mixed Membership Models*, In submission.
- H. Wang, H. Shan, A. Banerjee, *Bayesian Cluster Ensembles*, SDM, 2009.

- **Acknowledgements**
  - Research supported by NASA grant NNX08AC36A
  - My Students: Hanhuai Shan, Amrudin Agovic