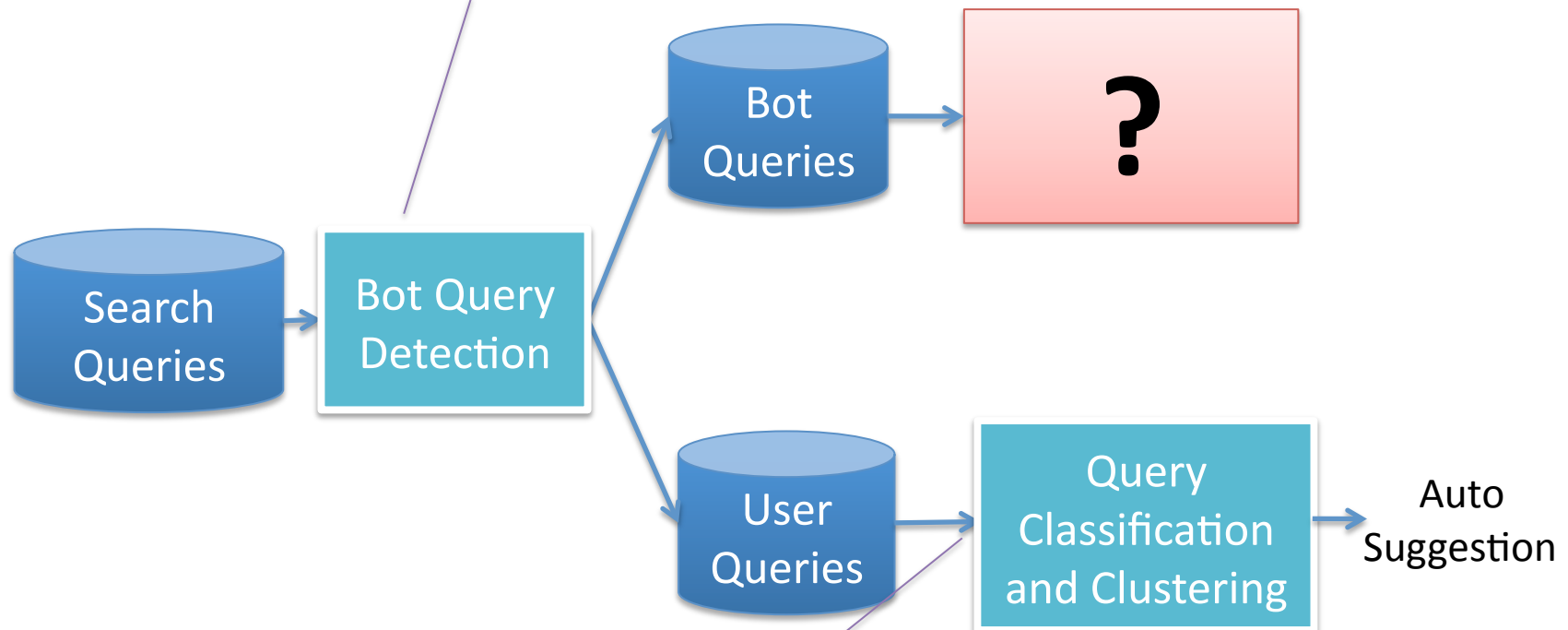


Motivations

- Huge volume of bot queries today
 - Finding security vulnerabilities [Usenix Sec 2010]
 - Search engine optimizations [Usenix Sec 2011]
 -
- **Understanding bot queries can be a new direction to tackle network security**
 - For security analysts: identify attack trends and detect botnets
 - For search engines: throttle malicious activities and protect good users

Related Work

H.Kang, et al., *“Large-Scale Bot Detection for Search Engines”*, WWW 2010
F. Yu, et al., *“Large-Scale Search Bot Detection”*, WSDM 2010
.....



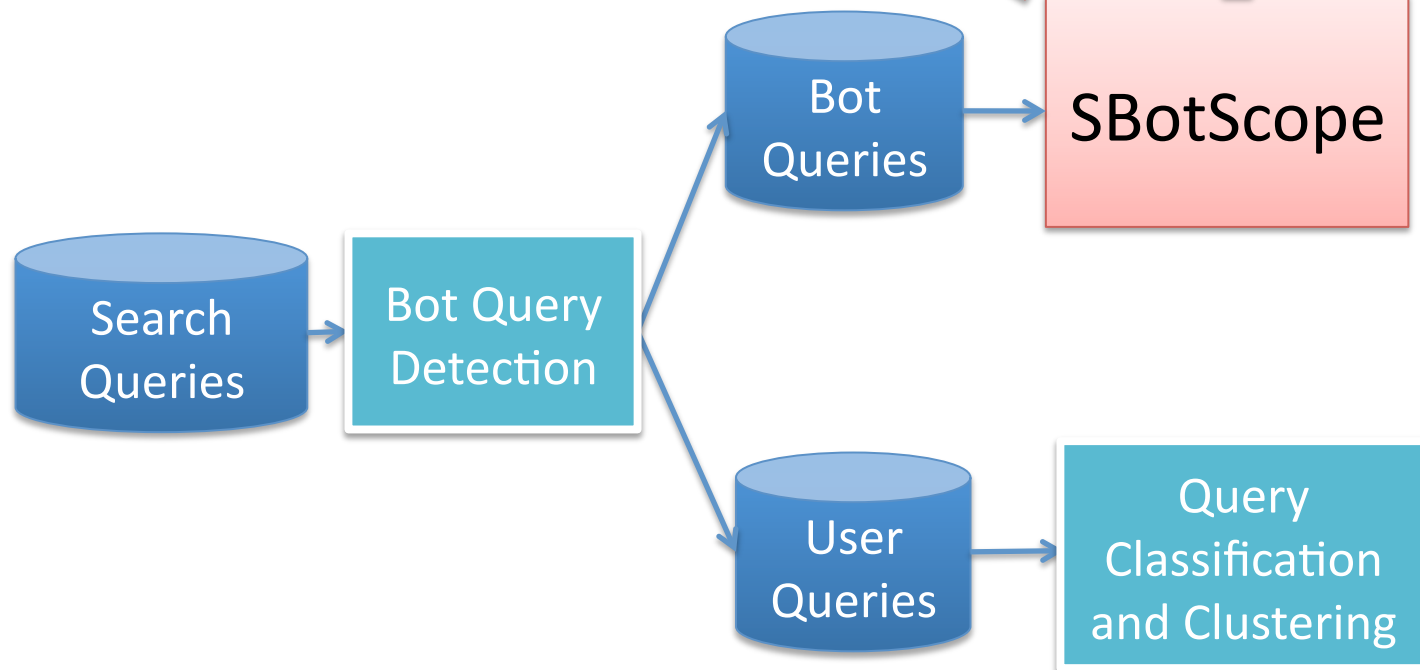
H. Cao, et al., *“Context-aware Query Classification”*, SIGIR09.

.....

Our Work: SBotScope

- What are bot queries searching for?

- Who submitted bot queries?
Botnets? Data centers?



Challenges

- Different behaviors compared to real users
 - Fast-changing topics
 - Few bot queries have clicks

- Various obfuscation strategies

- Mix truly intended queries with legitimate ones

114.46.61.111 5/21/2011 google

114.46.61.111 5/21/2011 lexingtn

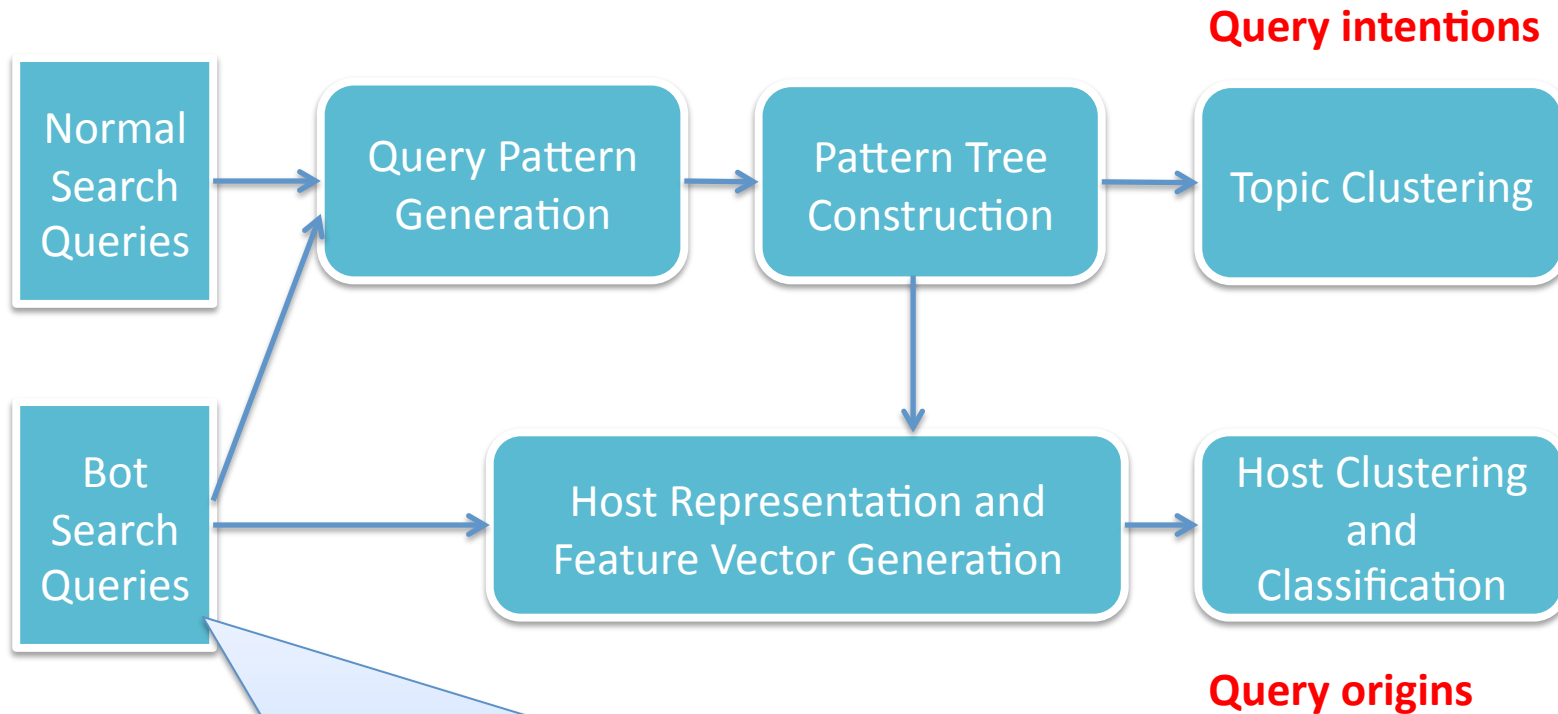
114.46.61.111 5/21/2011 Youtube

- Large data volume

- No data labeling

fort collins auto insurance BBCFER
fort collins auto insurance KYEFE
SDUEF fort collins auto insurance

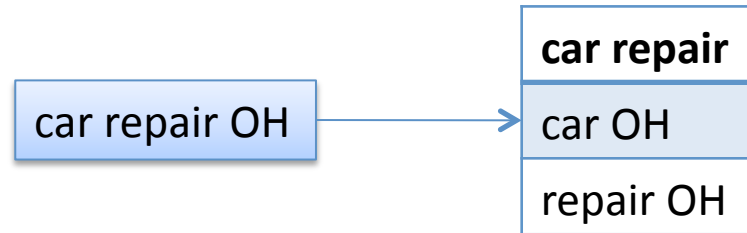
System Architecture



IP	Time	Query	UserAgent, FormCode,Click
12.23.x.x	2011/5/11/9-22-11	php powered by
23.12.x.x	2011/5/11/9-22-12	car insurance
.....			

Query Pattern Generation

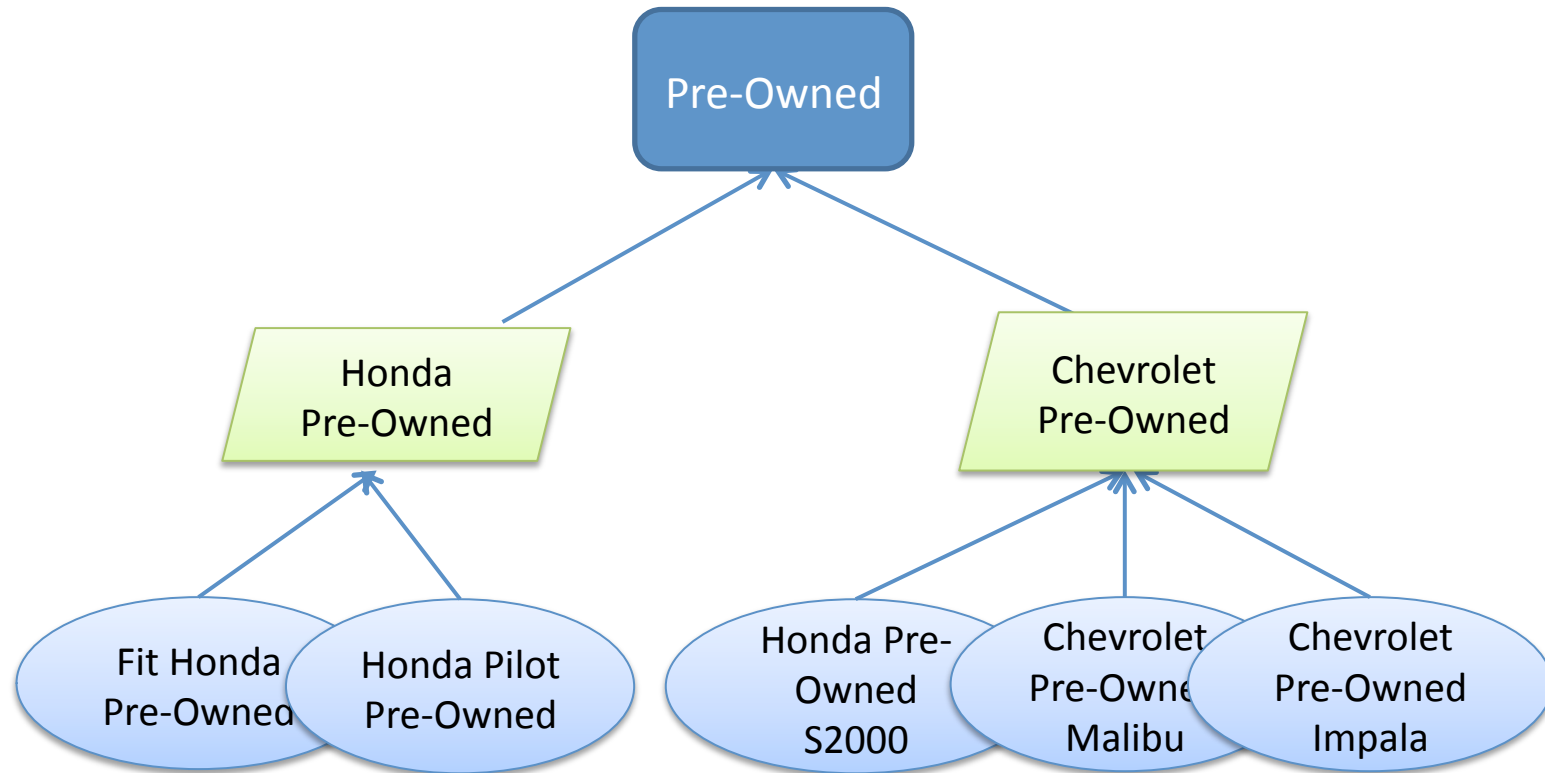
- Word Combinations
 - Robust to word order changes



- Pattern: a **specific** word-combination that happens **frequently**
 - Frequent: frequency rankings
 - Specific: conditional entropy + normal query popularity

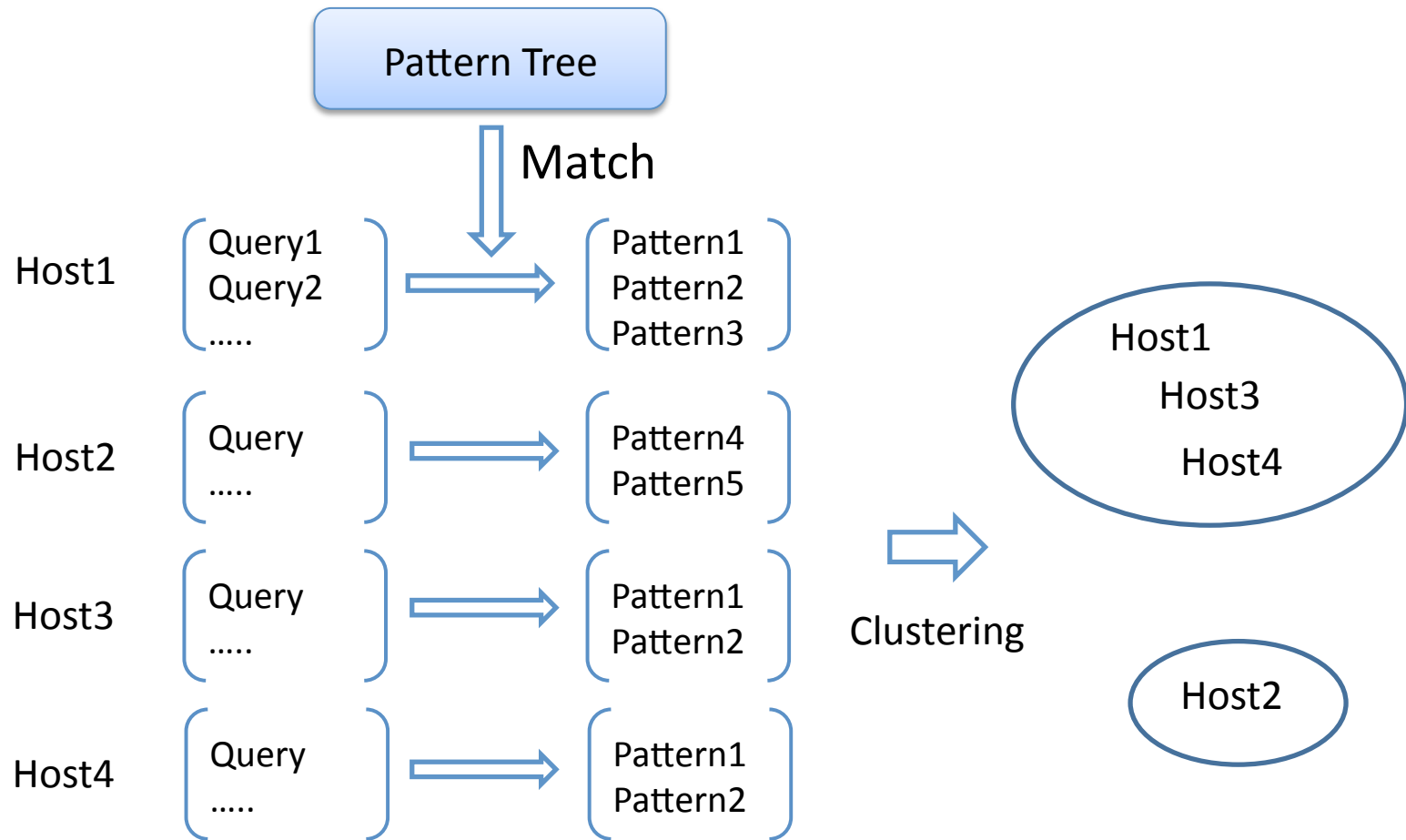
Topic Analysis

- **Syntactically:** construct pattern trees hierarchically

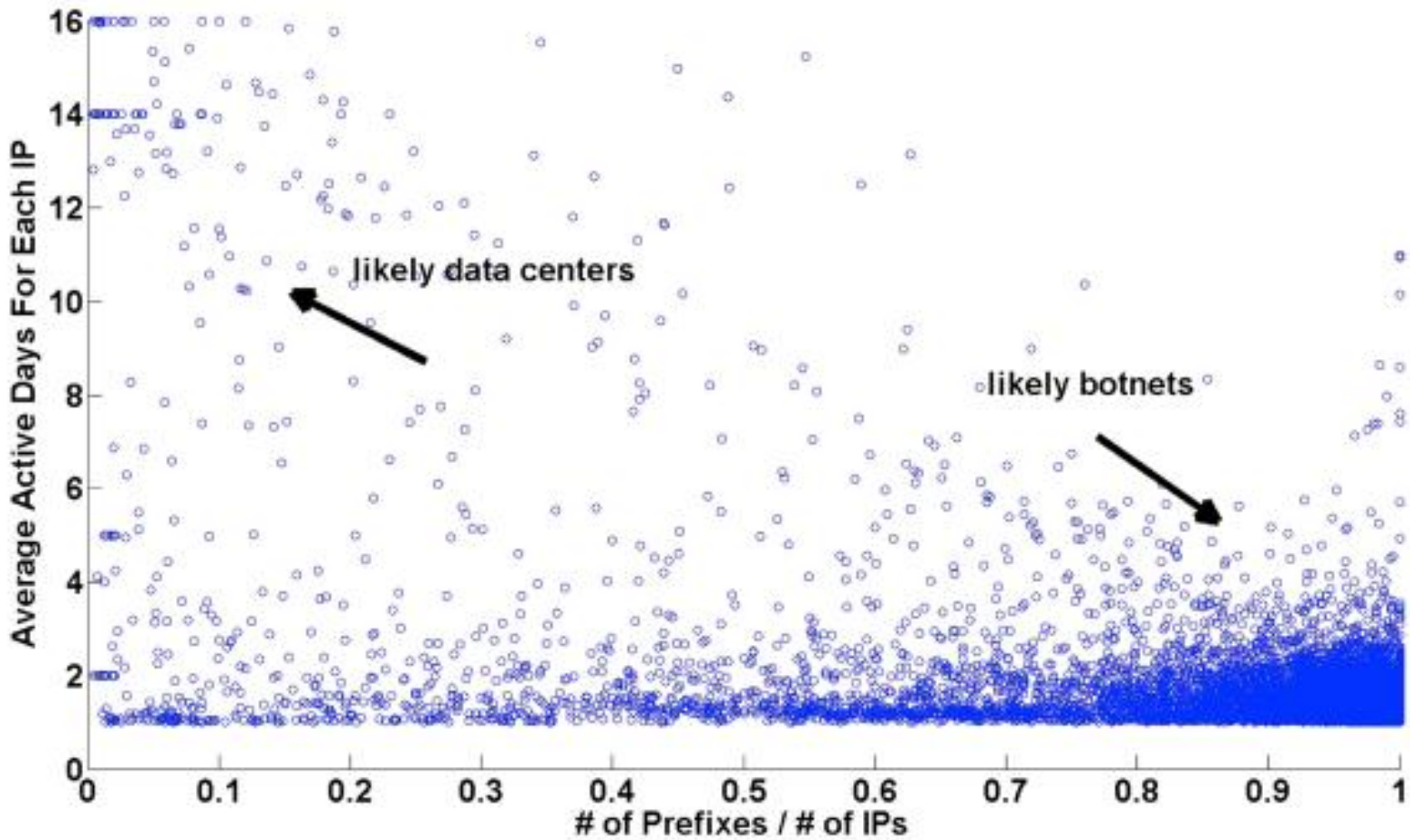


- **Semantically:** group trees into topics via spectral clustering

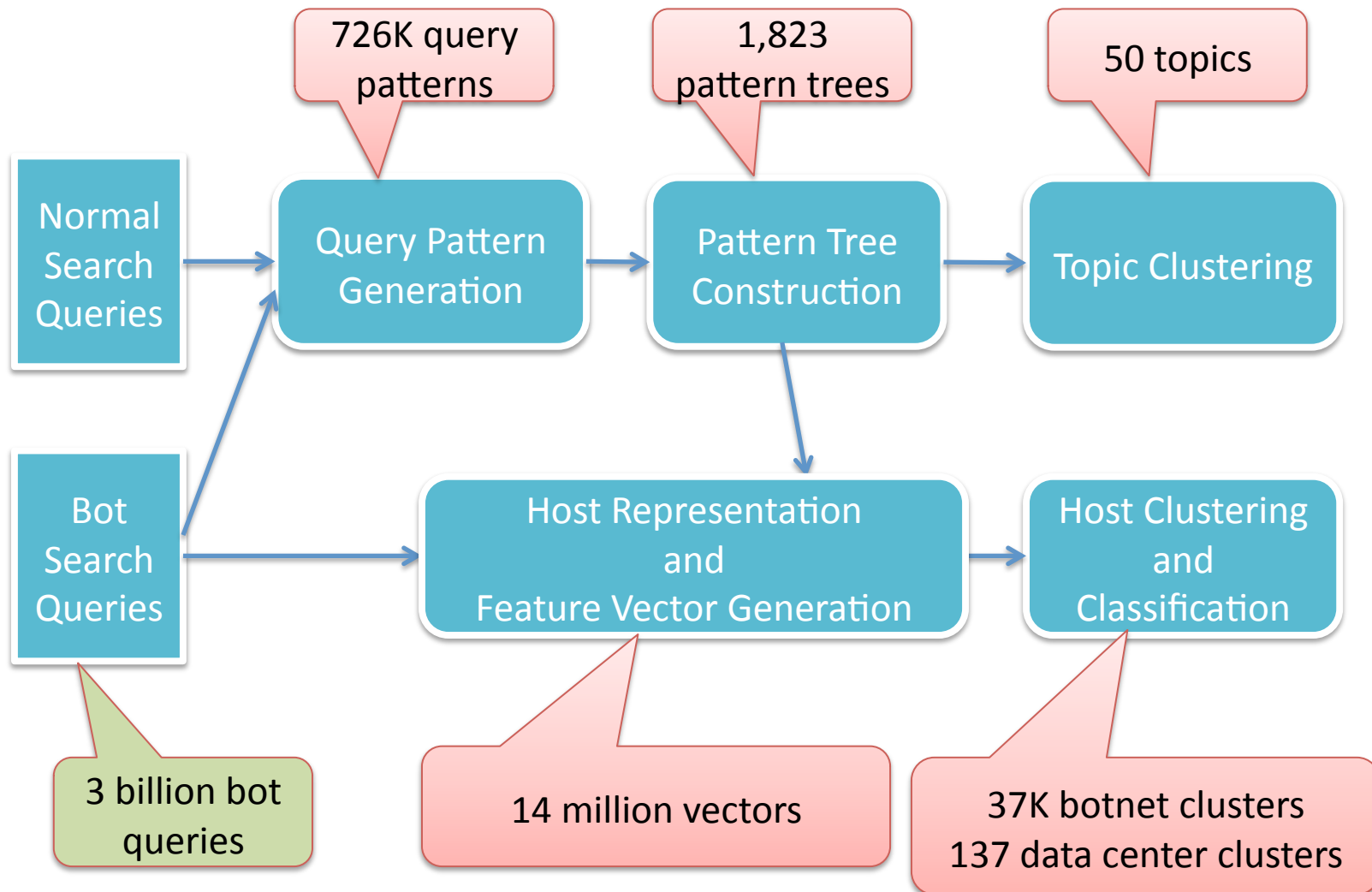
Feature Vector Generation and Host Clustering



Host Cluster Classification



Results: Complexity Reduction



Intentions: Popular Topics

Topic	% of Bot Queries
Vulnerability Discovery	32.8%
Email Harvest	11%
Content Download	3.6%
Fashion Items	1.4%
Car Sale	1.3%
News	0.7%
.....

Top 6 popular topics

Intentions: Popular Topics

Topic	% of Bot Queries
Vulnerability Discovery	32.8%
Email Harvest	11%
Content Download	3.6%
Fashion Items	
Car Sale	
News	
.....	

Top 5 Patterns
list members mode php mode php register es php page aspx html php powered by

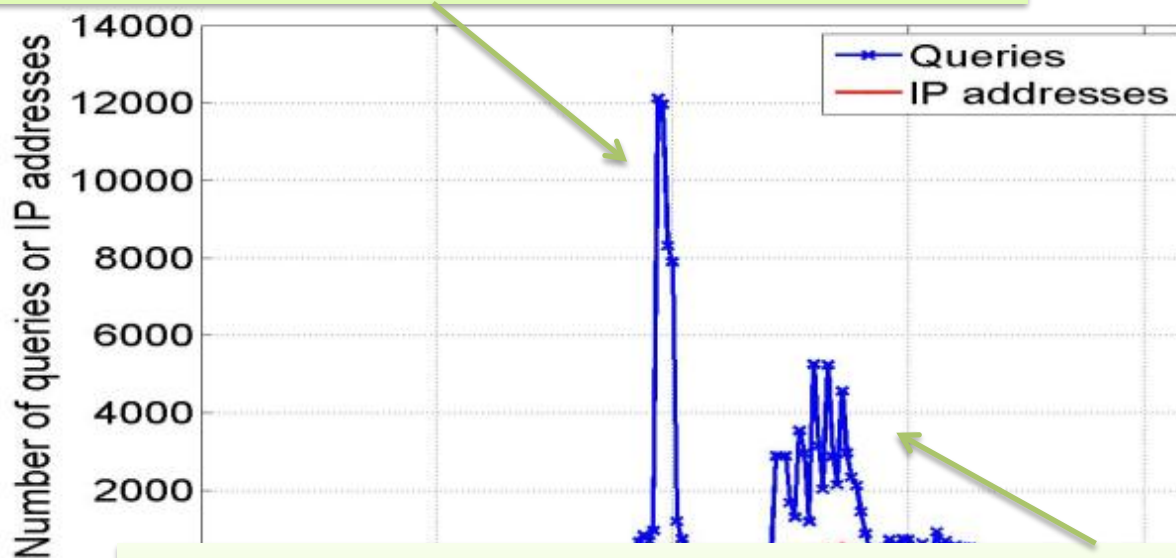
Query Patterns	Vulnerabilities
powered by photo album	PHP Album 0.3.2.3 Remote command execution
powered by update	PHP-Update 2.7 Remote code execution
Powered by icalender	PHP-iCalendar 2.24 File upload

Origination: Botnets vs. Data Centers

	Botnets	Data Centers
	php, download, powered, email, pdf	light, sold, video, bank, car
Number	37,268	137
Intentions	Vulnerability, email address	Commercially relevant information
Queries/host/day	17	475
Network distributions	Different organizations	The same organizations
Top country	CN, MX, IT, BR	US

Case Study I: A Botnet Cluster

“index.php/thread-” shooter
“yabb/yabb.pl?board” siver
“index.php/thread-” site:com
.....
act “WordPress forum plugin by Fredrik Fahlstad” site:.edu

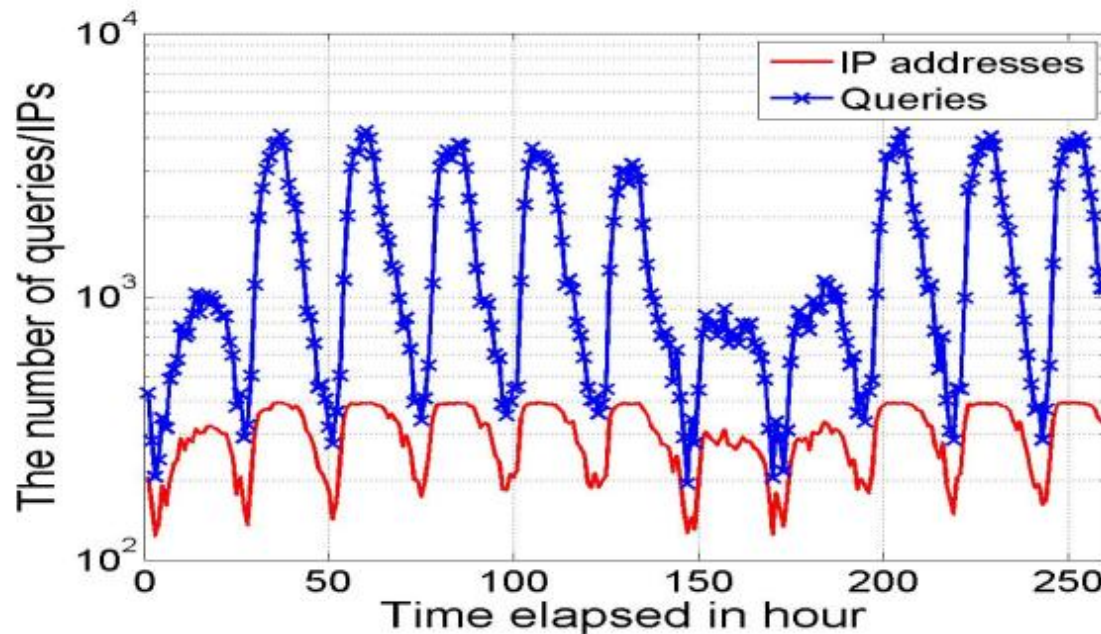


inch “WordPress forum plugin by Fredrik Fahlstad” site:.cn plus
“WordPress forum plugin by Fredrik Fahlstad” site:.net change
“WordPress forum plugin by Fredrik Fahlstad” site:.com

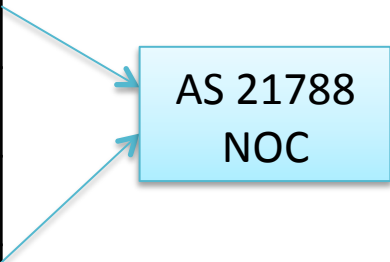
Case Study II: A Data Center Cluster

# of IPs	446
# of /24 prefixes	29
# of organizations	1
# of queries	873,064

“alice foo” + (“alice is” OR “alice was”) –Genealogy – Generation language:en



Case Study III: Vulnerability-Searching Data Centers

A cluster in May	# of IPs	41	
	# of organizations	1	
	# of queries	244,054	
A cluster in Oct	# of IPs	18	
	# of organizations	1	
	# of queries	113,994	

Query Pattern: “powered by”

Conclusion

- SBotScope allows systematic analysis of bot queries
 - Understand query intentions and origination
 - Increase detection coverage
 - Fully automated and scalable
- A new direction to improve network security
 - Identify attack trends at their early stages
 - Detect botnets and malicious data center hosts

Thank you!

- Questions ...