SKYNET: Applying Advanced Cloud-based Behavior Analytics

A Collaborative Project by S2I, R6, T12, T14, SSG, and S22

Presenters:

[redacted], S2I51

[redacted], R66F

Derived From: NSA/CSSM 1-52
Dated: 20070108
Declassify On: 20370401

Outline

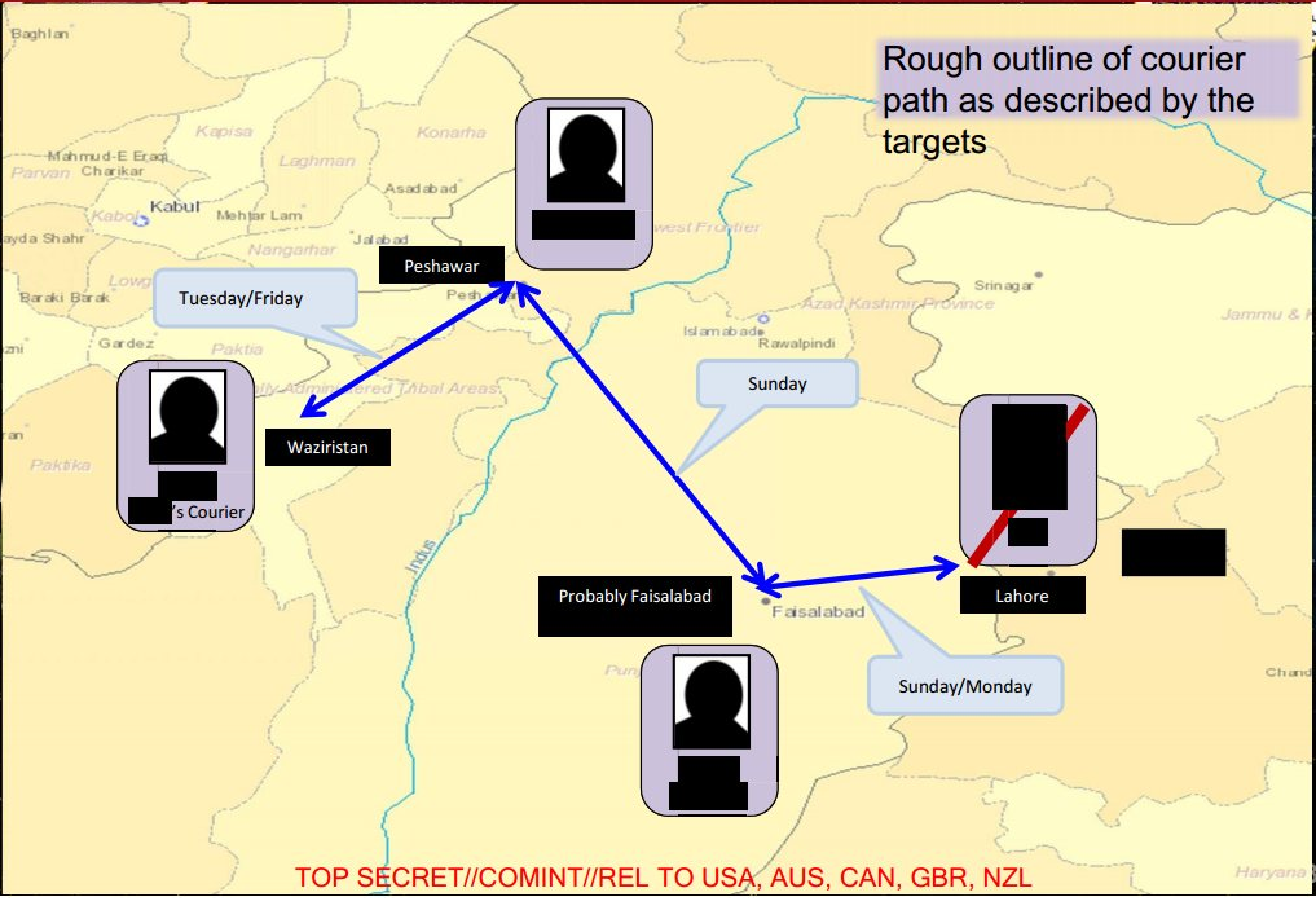
- What is SKYNET?
- DEMONSPIT Data Flow
- Automated Bulk Cloud Analytics
- Analytic Triage

What is SKYNET?

- Collaborative cloud research effort between 5 different organizations crossing 3 NSA Directorates:
  - Signals Intelligence: S2I, S22, SSG
  - Research: R6
  - Technology: T12, T14
- Partnerships
  - TMAC/FASTSCOPE
  - MIT Lincoln Labs & Harvard
- **SKYNET applies complex combinations of geospatial, geotemporal, pattern-of-life, and travel analytics to bulk DNR data to identify patterns of suspect activity**

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

Intelligence Update



Rough outline of courier path as described by the targets

[redacted]

Peshawar

Tuesday/Friday

Waziristan

[redacted]

[redacted]'s Courier

Sunday

[redacted]

[redacted]

[redacted]

Lahore

Probably Faisalabad

[redacted]

Sunday/Monday

[redacted]

TOP SECRET//COMINT//REL TO USA, AUS, CAN, GBR, NZL

SKYNET Analytic Questions

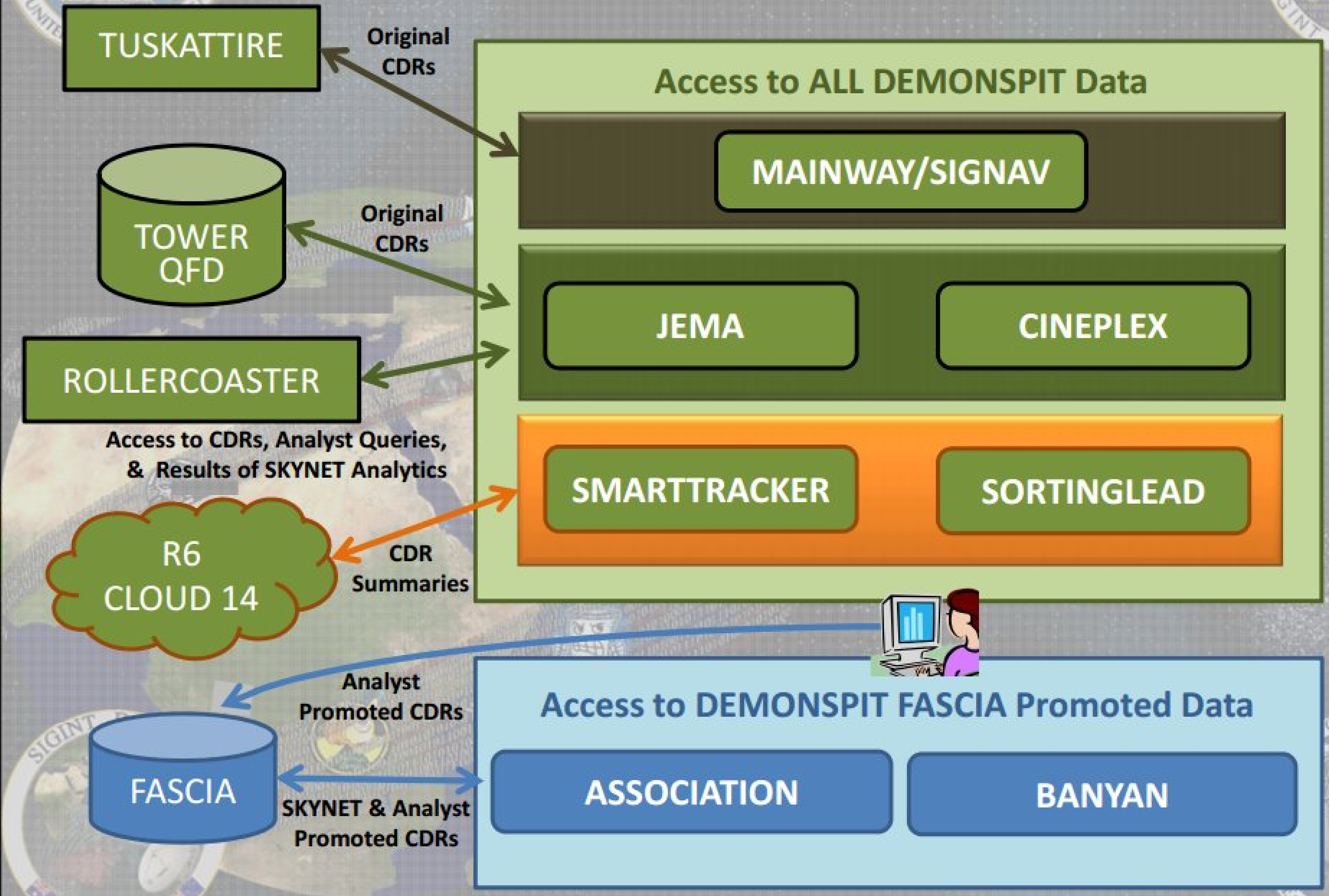
- Who has traveled from Peshawar to Faisalabad or Lahore (and back) in the past month?
  - Who does the traveler call when he arrives?
  - Who else is seen in the area when the traveler arrives, and who seen leaving the area shortly afterward?
- Who travels to/from Peshawar every other Sunday and "somewhere else“ on a weekly basis?
- Who visits Akora Khattak periodically and also travels between Peshawar and Lahore?
- Who fits the above travel profiles and also possesses unusual behavior:
  - One or two hops from other suspects or known tasked selectors
  - Frequent handset swapping or powering down

DEMONSPIT

- DEMONSPIT is a new dataflow for bulk Call Data Records (CDRs) from Pakistan
  - CDRs are being acquired from major PK Telecom providers
- Data is normalized through TUSKATTIRE, like all other Call Data Records
- DEMONSPIT data is forwarded by TUSKATTIRE to several Clouds:
  - GMHalo/DPS
    - Promotes records to FASCIA and feeds the SEDB Tower QFD
  - GMPlace & Cloud 14
    - Ingests DEMONSPIT into Sortinglead summaries to support SKYNET Analytics
    - Ingests DEMONSPIT into a Perishable QFD which will be available to analysts via JEMA and CINEPLEX
  - Bulldozer/MDR2

All of the clouds receiving DEMONSPIT data also receive all FASCIA data

Analysts' View of DEMONSPIT



Outline

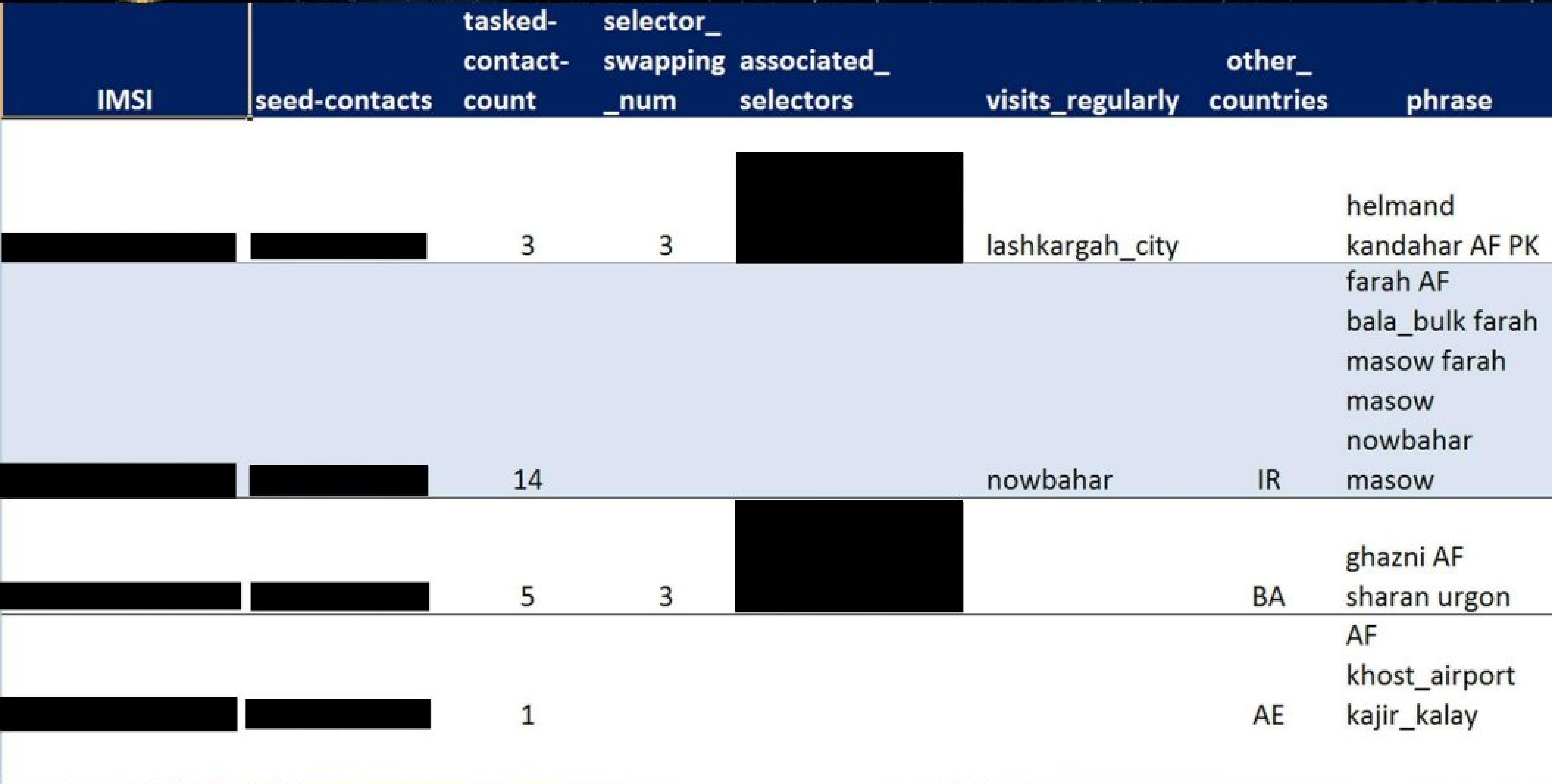
- What is SKYNET?
- DEMONSPIT Data Flow
- Automated Bulk Cloud Analytics
- Analytic Triage

[redacted]

Cloud Analytic Building Blocks

- Travel Patterns
  - Travel phrases (Locations visited in given timeframe)
  - Regular/repeated visits to locations of interest
- Behavior-Based Analytics
  - Low use, incoming calls only
  - Excessive SIM or Handset swapping
  - Frequent Detach/Power-down
  - Courier machine learning models
- Other Enrichments
  - Travel on particular days of the week
  - Co-travelers
  - Similar travel patterns
  - Common contacts
  - Visits to airports
  - Other countries
  - Overnight trips
  - Permanent move

Sample Travel Report: Haqqani Network



IMSI	seed-contacts	tasked-contact-count	selector_swapping_num	associated_selectors	visits_regularly	other_countries	phrase
[redacted]	[redacted]	3	3	[redacted]	lashkargah_city		helmand kandahar AF PK
[redacted]	[redacted]	14			nowbahar	IR	farah AF bala_bulk farah masow farah masow nowbahar masow
[redacted]	[redacted]	5	3	[redacted]		BA	ghazni AF sharan urgon
[redacted]	[redacted]	1				AE	AF khost_airport kajir_kalay

What Suspicious Selectors Were Seen Traveling Between Peshawar and Lahore?

Case-Specific Behavioral Cloud Analytics

Peshawar-Lahore Travel 1 - 4 NOV 2011

TRAVEL PHRASE	DOW	MSISDN	IMSI	TASKED CONTACTS	NUM_SELECTOR_SWAPPING	ASSOCIATED_SELECTORS	ACTIVITY_CATEGORIES
torkham AF PK peshawar lahore	FRI	[redacted]		2			
PK peshawar lahore	THU	[redacted]					
behsud AF jalalabad jalal_abad jalalabad behsud rodat bati_kot mohmand_darah peshawar PK	WED		[redacted]	4	7	[redacted]	
gtrd PK nowshera gulbahar peshawar sanda_kalan lahore	THU	[redacted]					
jamrud PK peshawar lahore	TUE	[redacted]		10			
PK peshawar lahore	THU		[redacted]				5-or-fewer-contacts, sms-and-zero-duration-calls-only, low-use

Outline

- What is SKYNET?
- DEMONSPIT Data Flow
- Automated Bulk Cloud Analytics
- Analytic triage
  - SMARTTRACKER
  - RT-RG
  - JEMA

[redacted]

Selectors of Interest from Cloud Travel Analytic

[redacted] (tasked)

IMSIs:

[redacted]

Handsets:

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

[redacted]

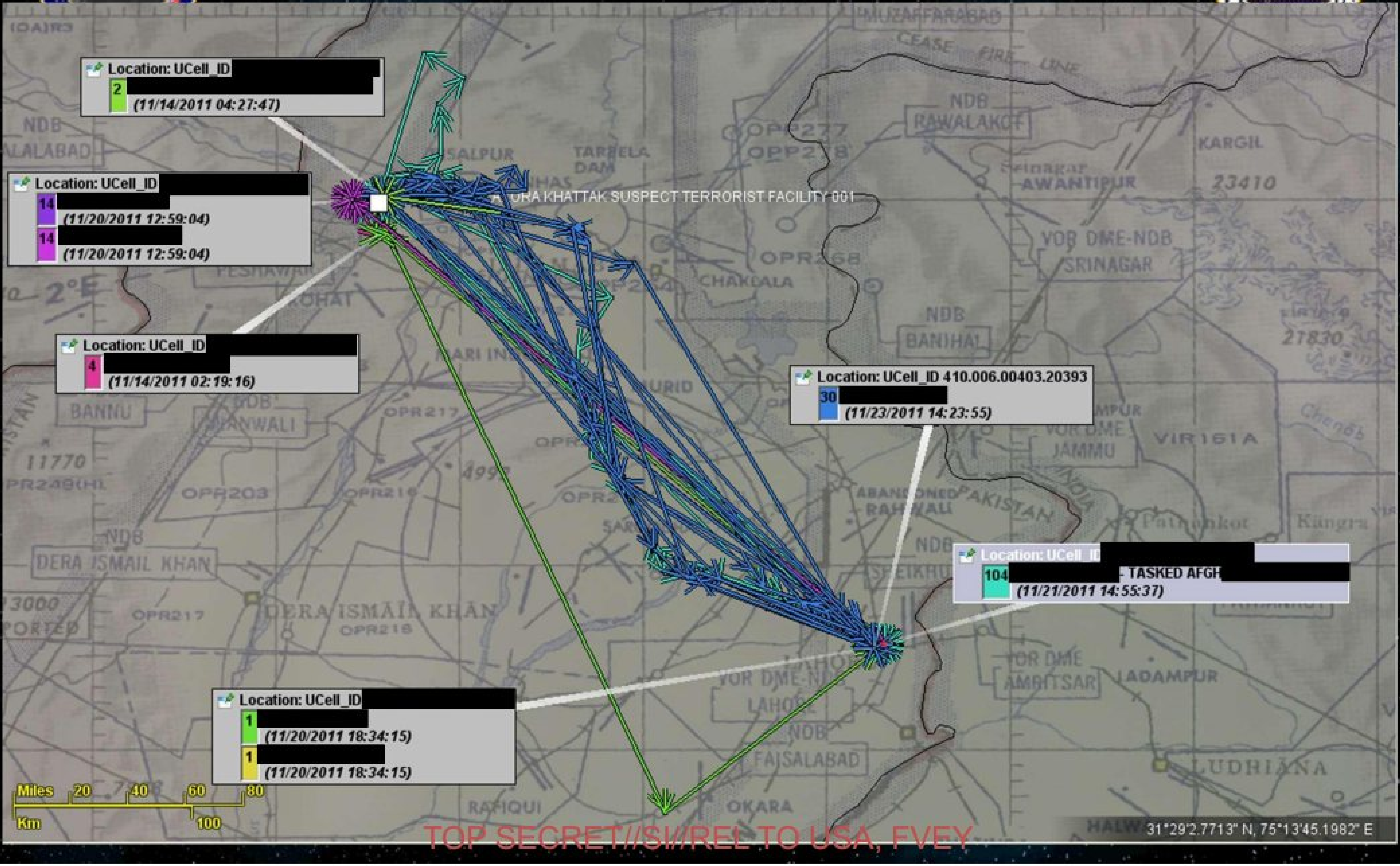
[redacted]

[redacted]

[redacted]

SMARTTRACKER Travel View

31 October – 23 November



Location: UCell_ID [redacted]
2 [redacted]
(11/14/2011 04:27:47)

Location: UCell_ID [redacted]
14 [redacted]
(11/20/2011 12:59:04)
14 [redacted]
(11/20/2011 12:59:04)

Location: UCell_ID [redacted]
4 [redacted]
(11/14/2011 02:19:16)

AKORA KHATTAK SUSPECT TERRORIST FACILITY 001

Location: UCell_ID 410.006.00403.20393
30 [redacted]
(11/23/2011 14:23:55)

Location: UCell_ID [redacted]
104 [redacted] - TASKED AFGH [redacted]
(11/21/2011 14:55:37)

Location: UCell_ID [redacted]
1 [redacted]
(11/20/2011 18:34:15)
1 [redacted]
(11/20/2011 18:34:15)

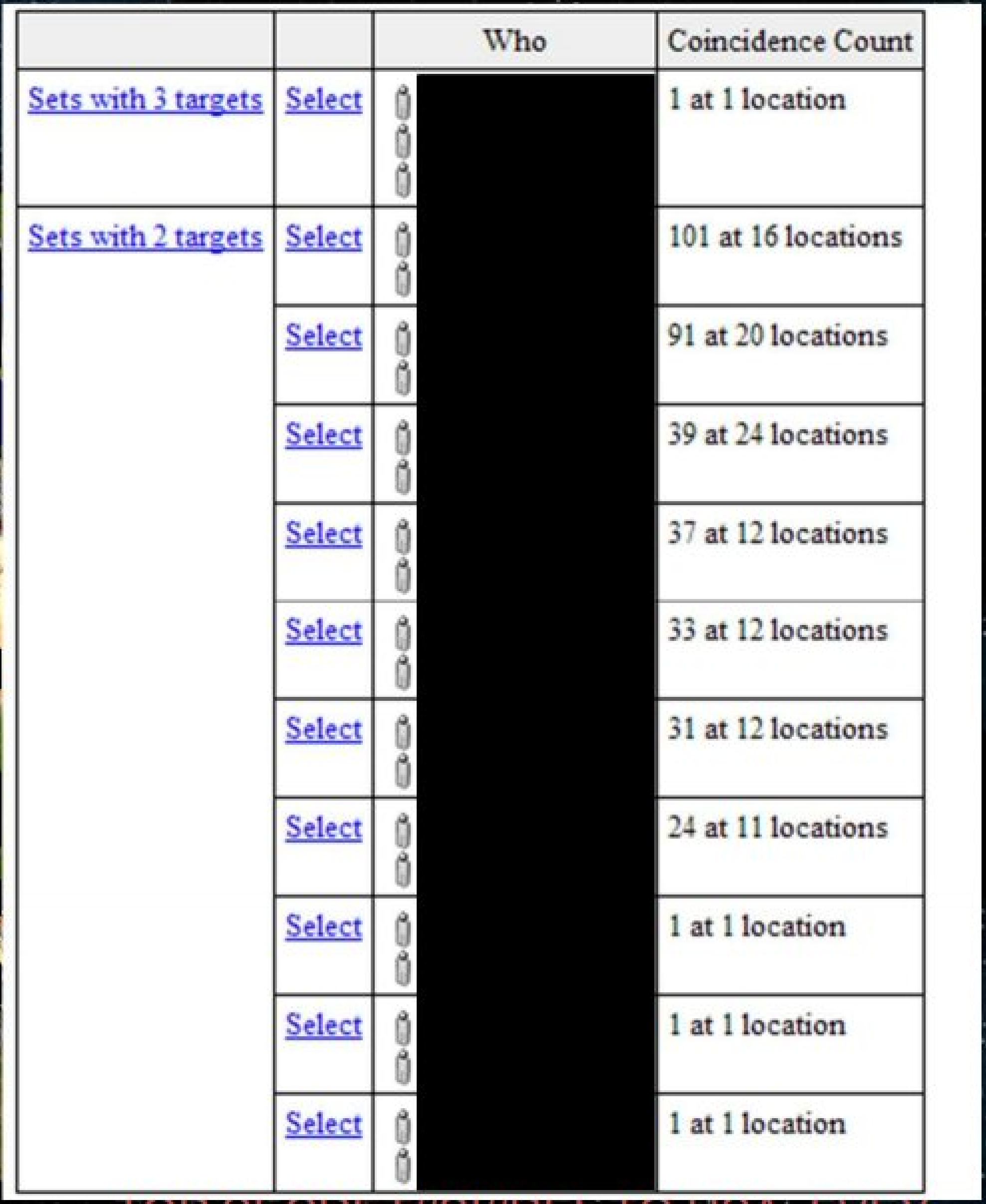
Analytic Tradecraft

- Examine travel patterns for common routes and meeting locations
  - Run cell soaks on all common meeting locations during meeting timeframe
- Analyze selectors for common contacts
- Analyze selectors for handset sharing behavior

Repeat procedure with resulting selectors

Correlate with other known and suspected selectors

SMARTTRACKER Coincidence Report



		Who	Coincidence Count
Sets with 3 targets	Select	[redacted]	1 at 1 location
Sets with 2 targets	Select	[redacted]	101 at 16 locations
	Select	[redacted]	91 at 20 locations
	Select	[redacted]	39 at 24 locations
	Select	[redacted]	37 at 12 locations
	Select	[redacted]	33 at 12 locations
	Select	[redacted]	31 at 12 locations
	Select	[redacted]	24 at 11 locations
	Select	[redacted]	1 at 1 location
	Select	[redacted]	1 at 1 location
	Select	[redacted]	1 at 1 location

[redacted]

[redacted]

[redacted]

[redacted]

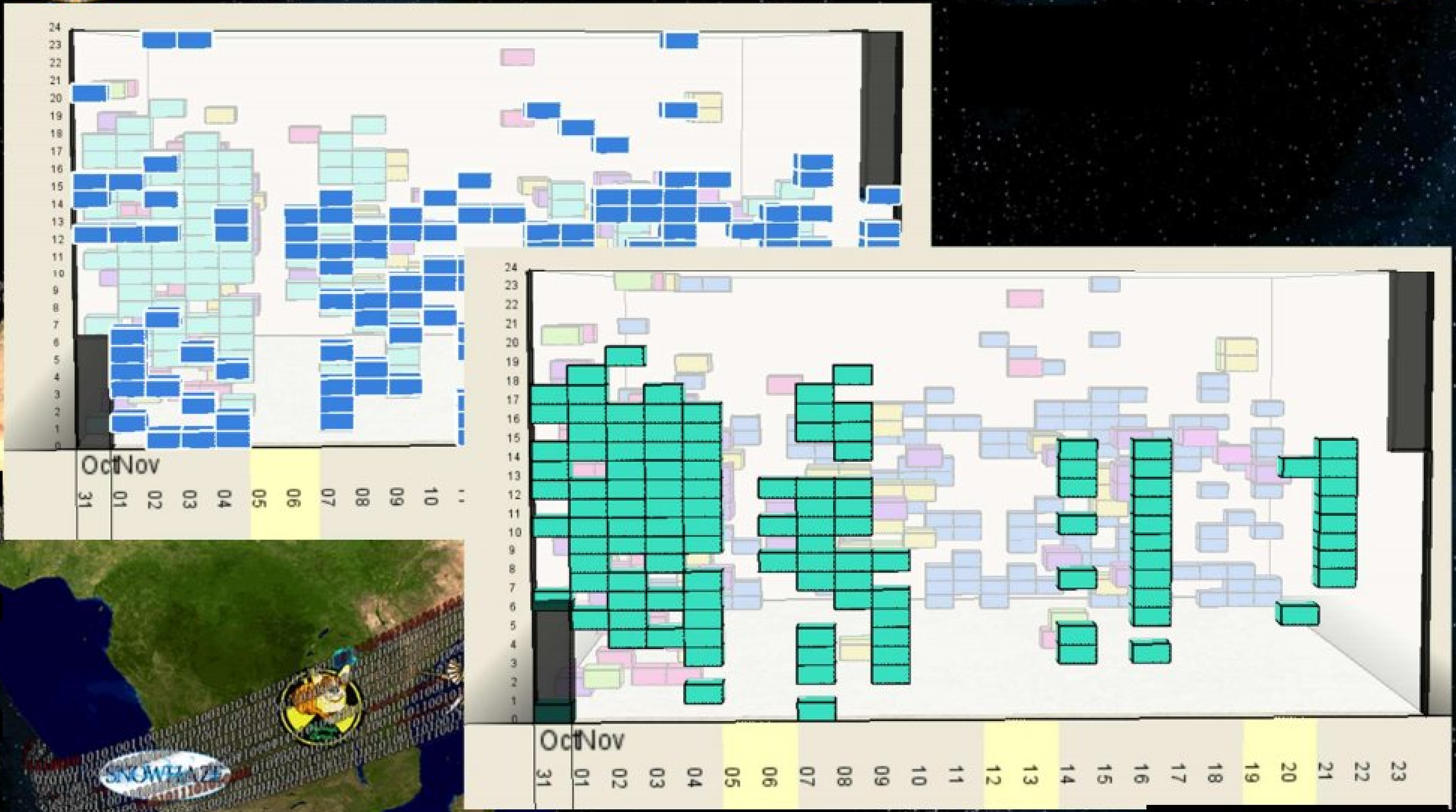
[redacted]

[redacted]

[redacted]

SMARTTRACKER

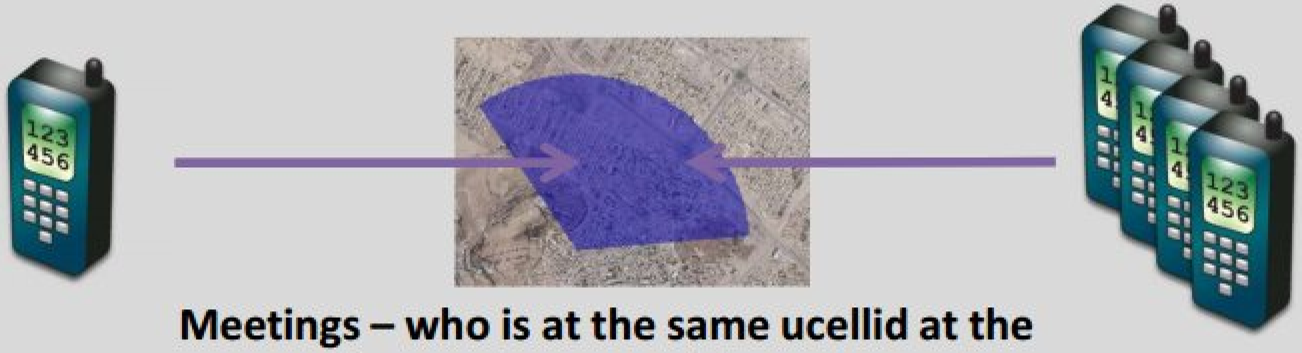
Smart Chart



[redacted]

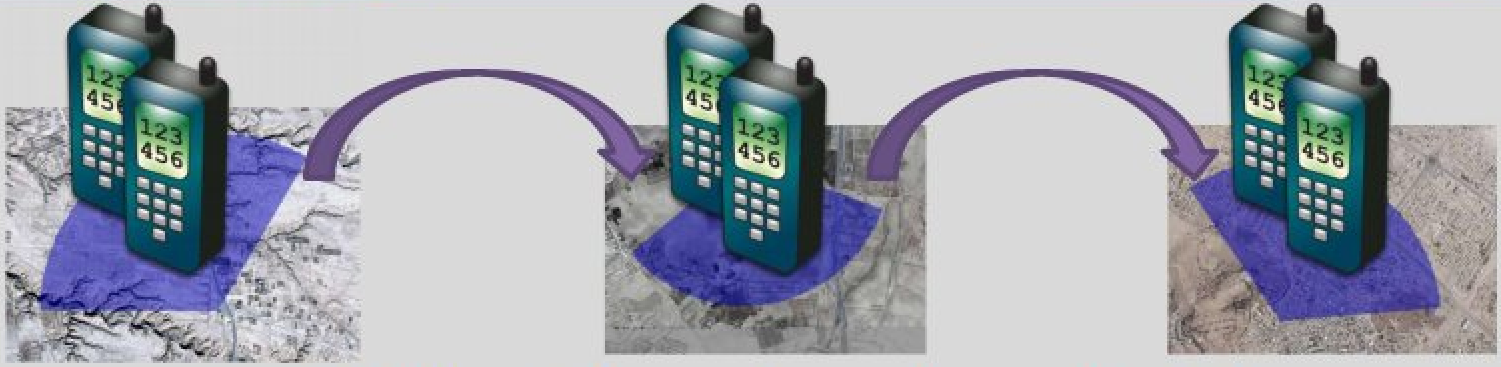
[redacted]

RT-RG Analytics



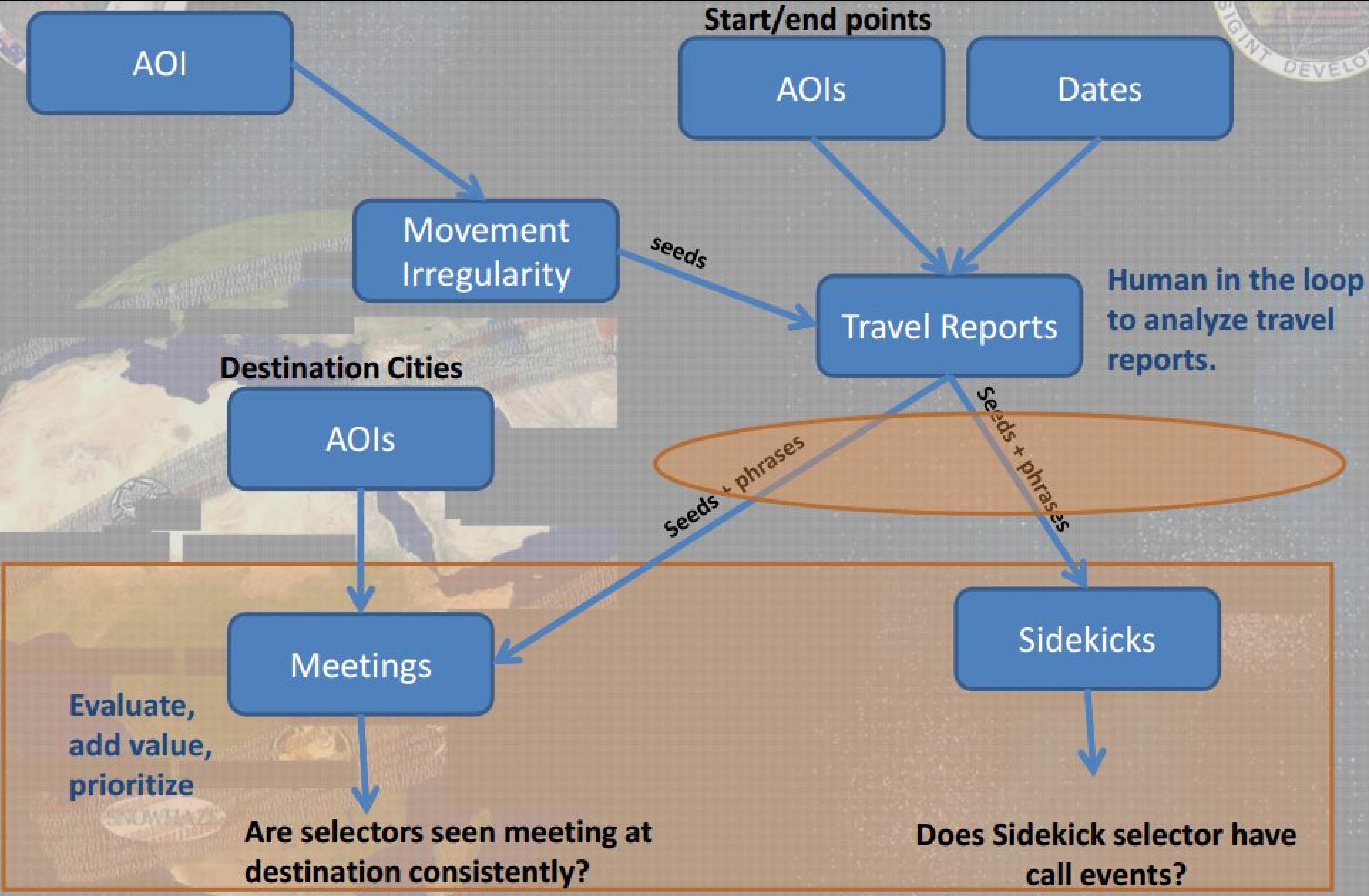
Meetings – who is at the same ucellid at the same time as the potential courier at the destination city?...Multiple times.

[redacted]



Sidekicks – is there a pair traveling together to the destination city?

JEMA: Pulling It All Together



AOI

Movement Irregularity

Start/end points

AOIs

Dates

Travel Reports

Human in the loop to analyze travel reports.

Destination Cities

AOIs

Meetings

Sidekicks

Evaluate, add value, prioritize

Are selectors seen meeting at destination consistently?

Does Sidekick selector have call events?

THANK YOU!

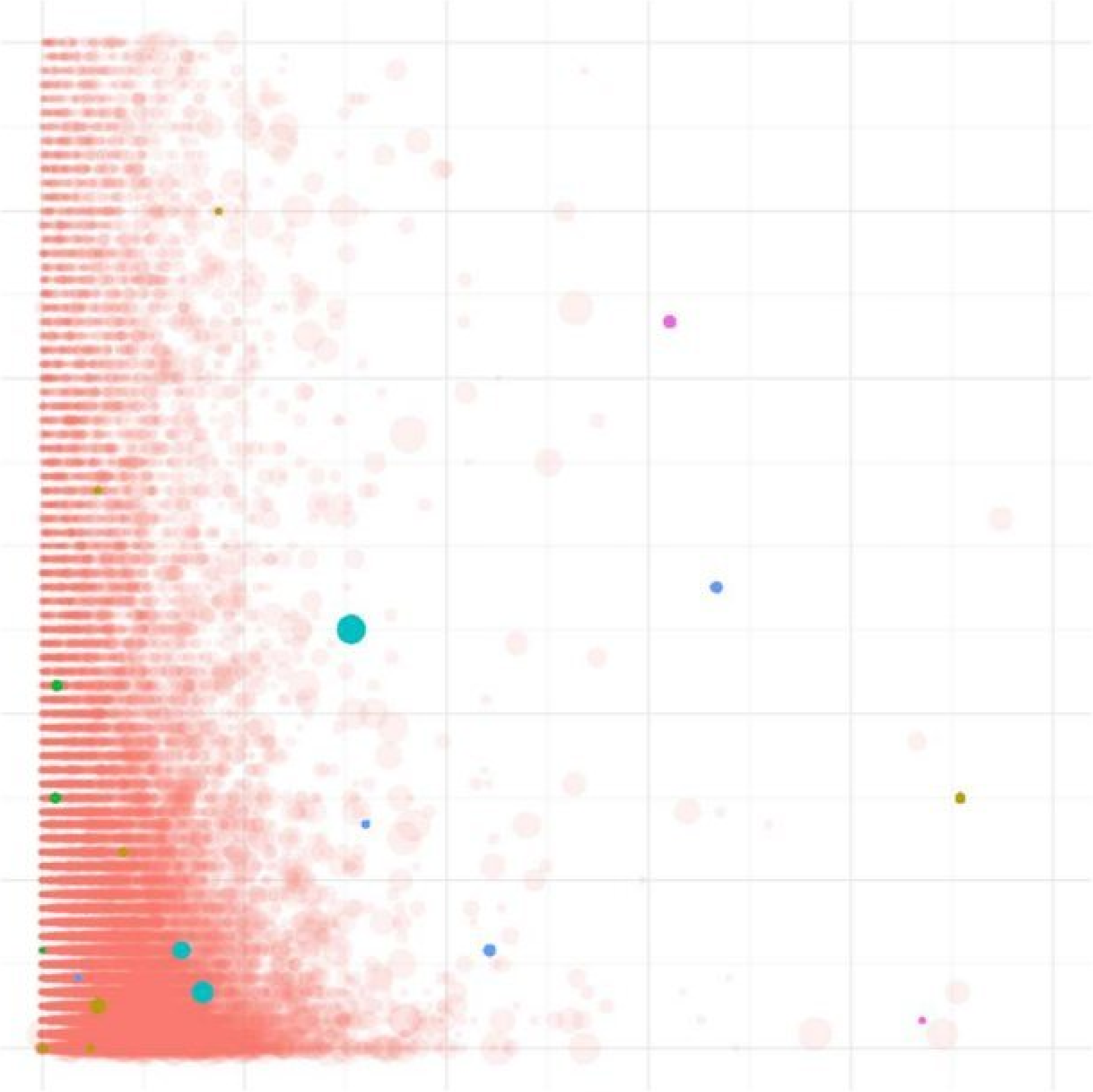
[redacted]

SKYNET WIKI:

https://[redacted]/wiki/SKYNET

[redacted], S2I51, [redacted]@nsa.ic.gov
[redacted], R66F, [redacted]@nsa.ic.gov

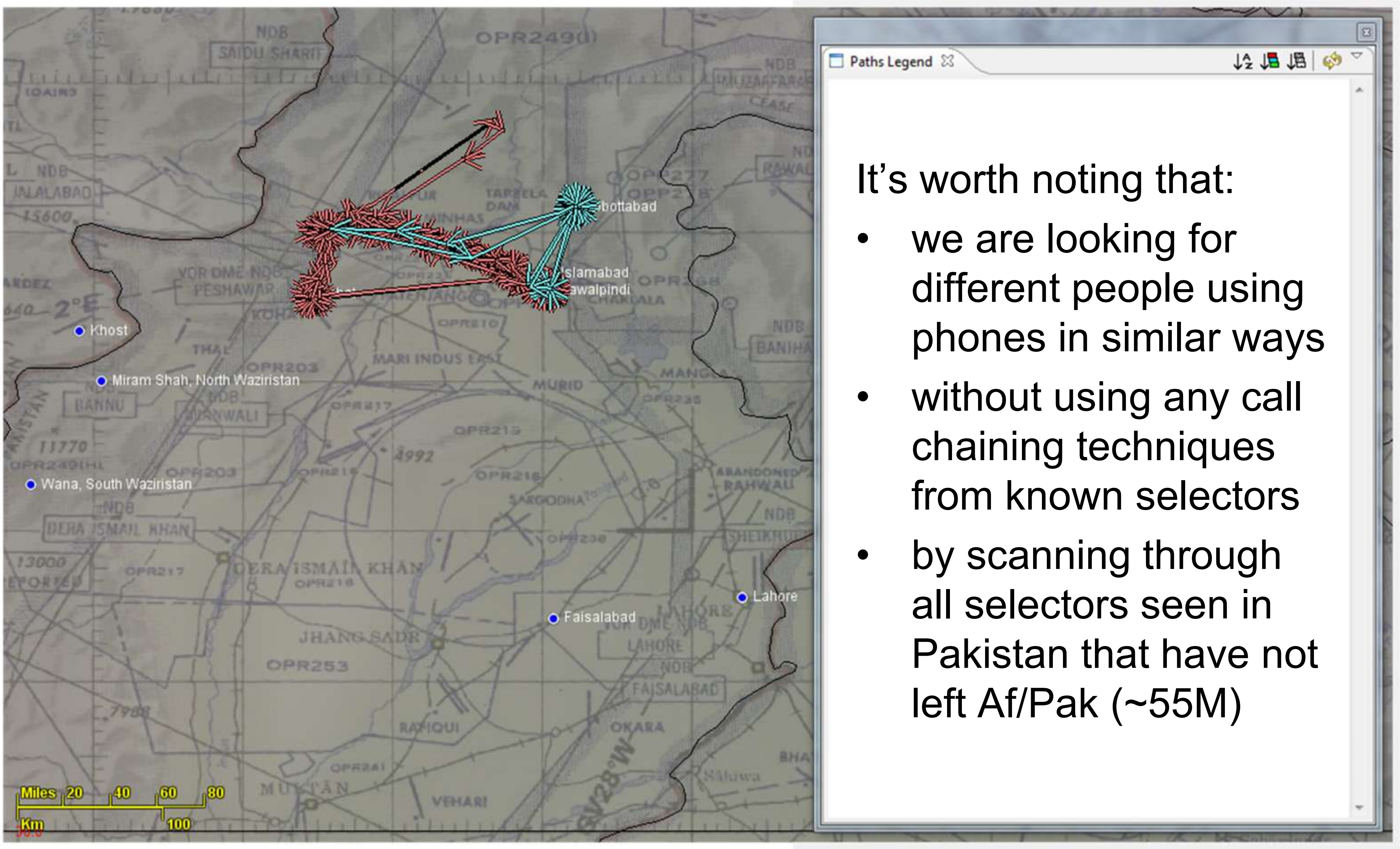
SKYNET:
Courier Detection via Machine Learning



[redacted], R66F/JHU
[redacted], R66F
[redacted], R66F
[redacted], T1211
[redacted], T1211
[redacted], S2I51
[redacted], S2I5/TD

June 5, 2012

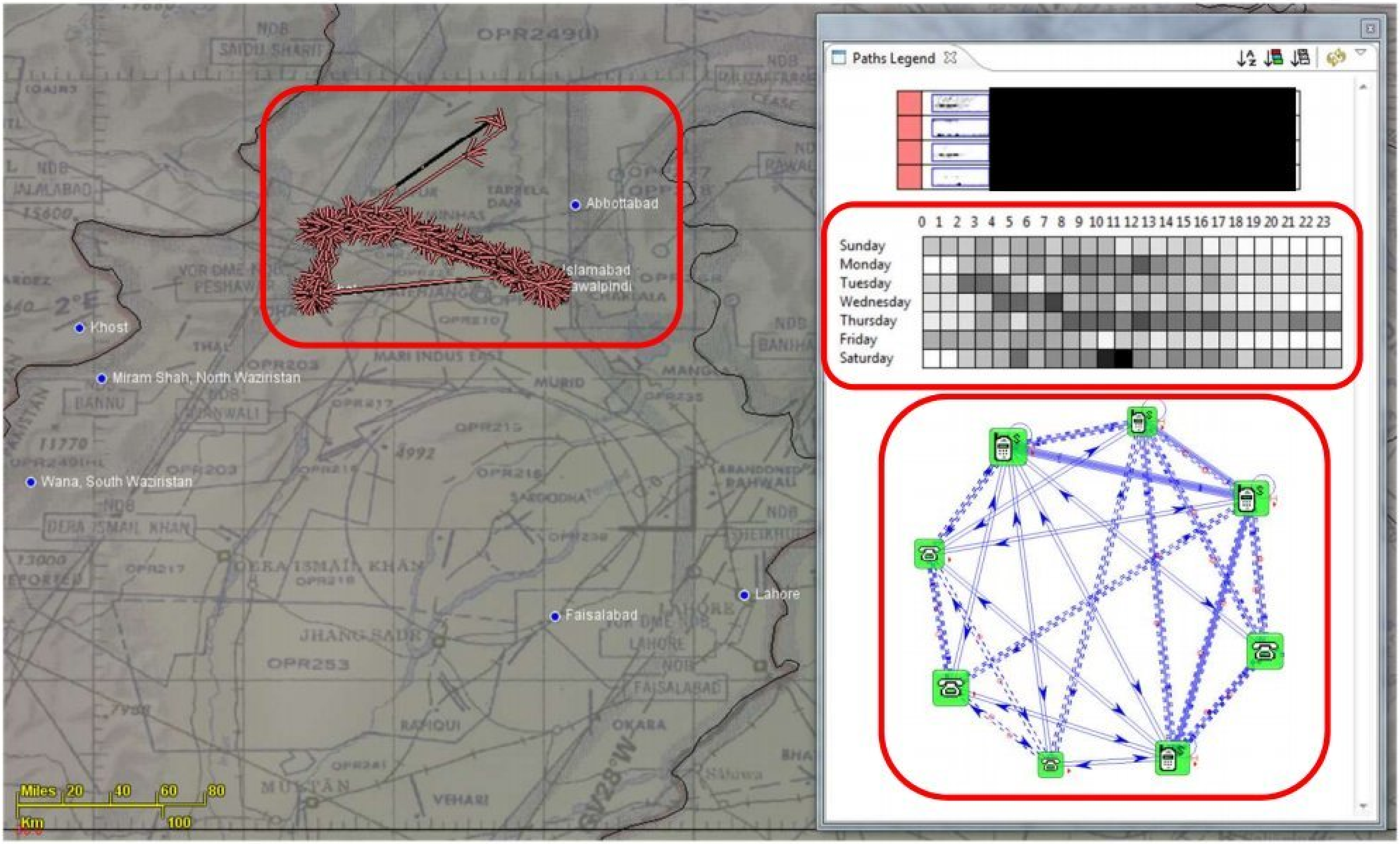
Given a handful of courier selectors, can we find others that "behave similarly" by analyzing GSM metadata?



It's worth noting that:

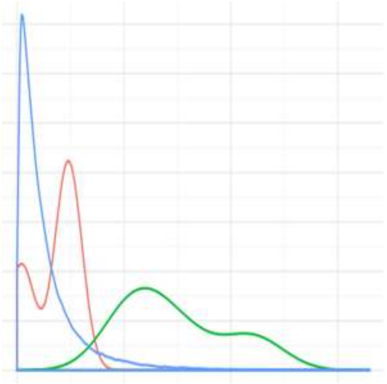
- we are looking for different people using phones in similar ways
- without using any call chaining techniques from known selectors
- by scanning through all selectors seen in Pakistan that have not left Af/Pak (~55M)

From GSM metadata, we can measure aspects of each selector's pattern-of-life, social network, and travel behavior

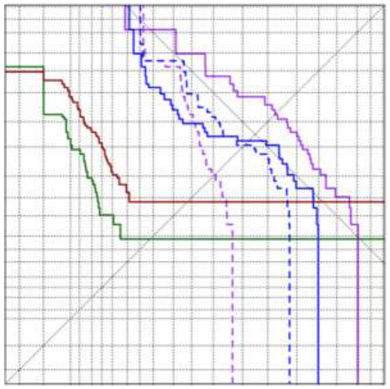


[redacted]

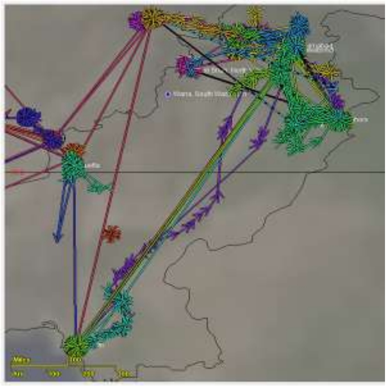
This presentation describes our search for AQSL couriers using behavioral profiling



Behavioral Feature Extraction

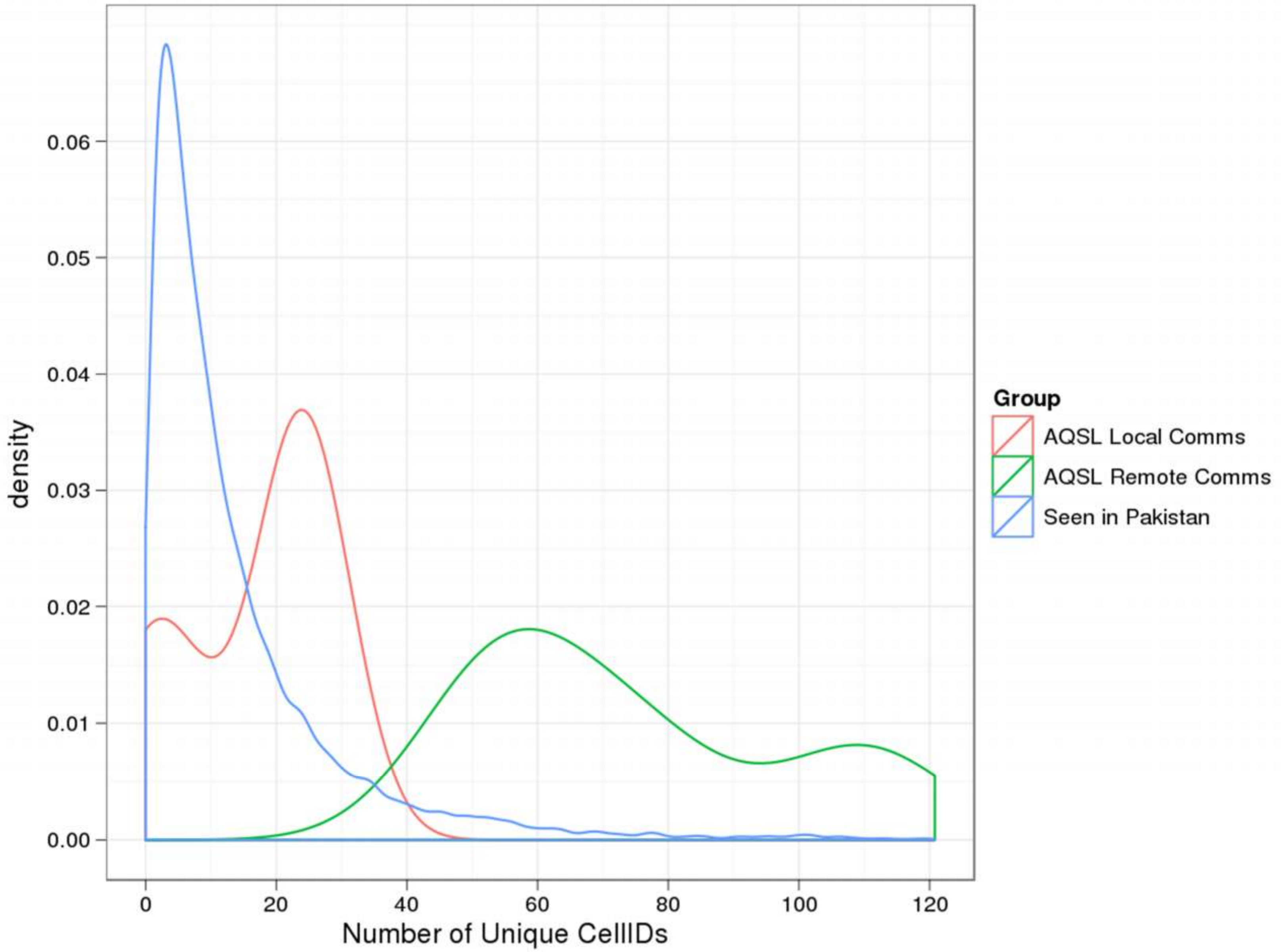


Cross Validation Experiment
on AQSL Couriers

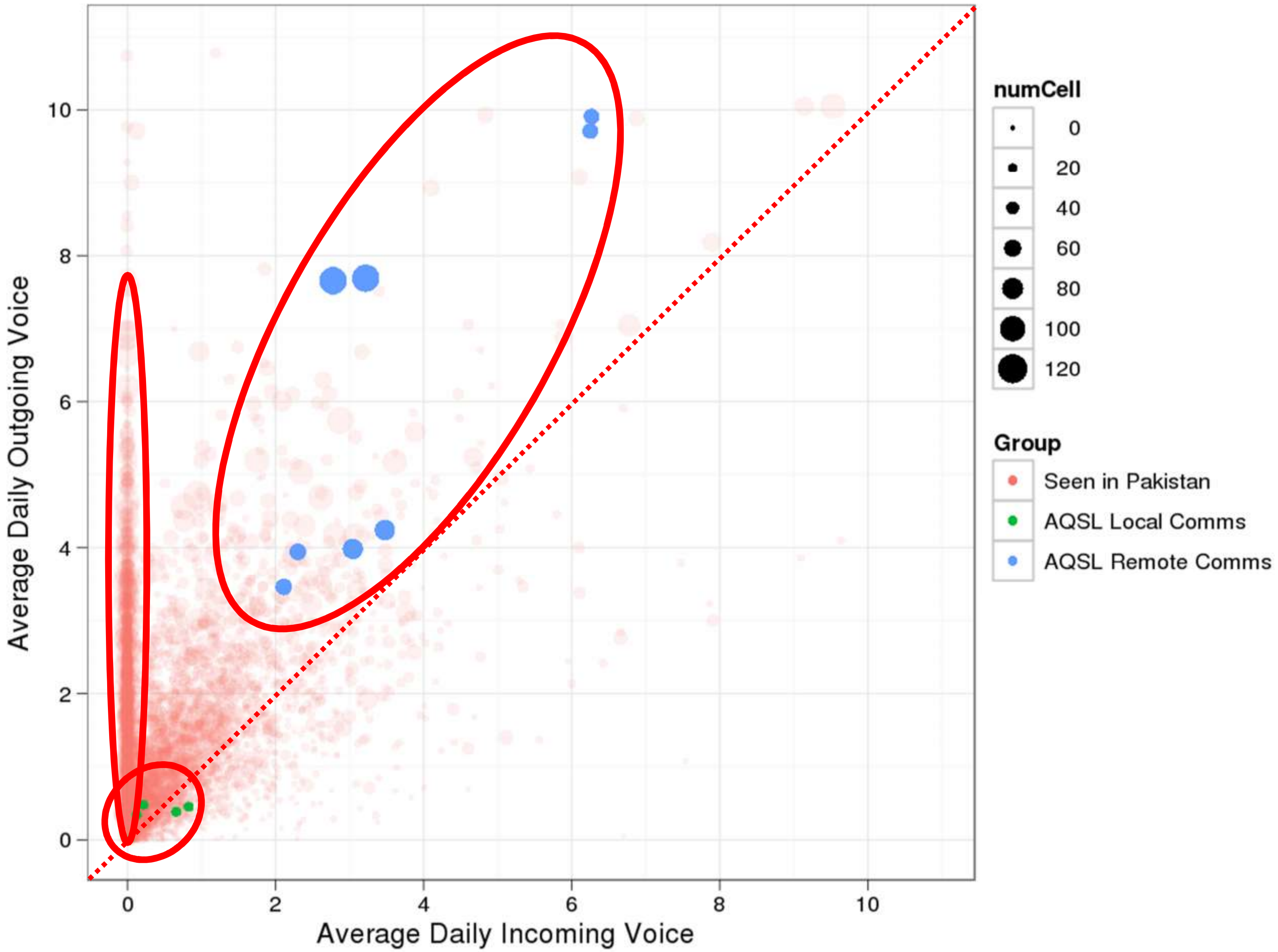


Preliminary SIGINT Findings

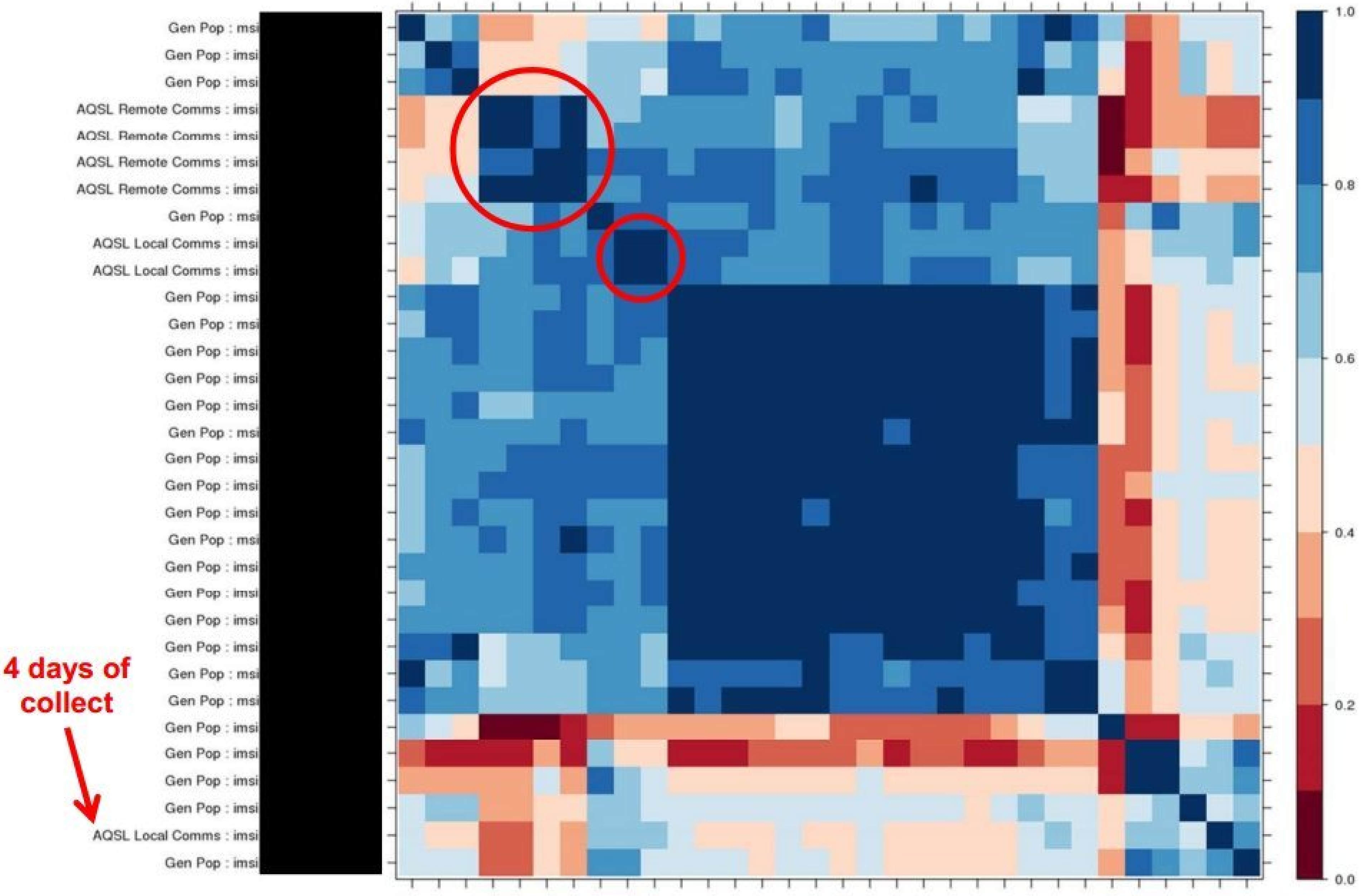
Counting unique UCELLIDs shows that couriers travel more often than typical Pakistani selectors



By examining multiple features at once, we can see some indicative behaviors of our courier selectors



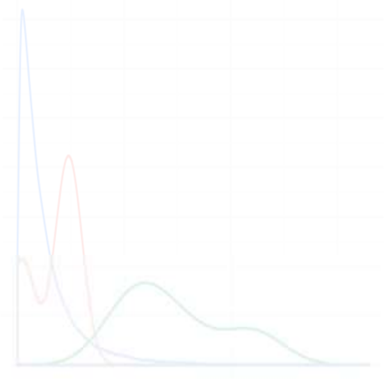
Looking at a hierarchical clustering derived from all 80 features, the AQSL groups mostly stay together



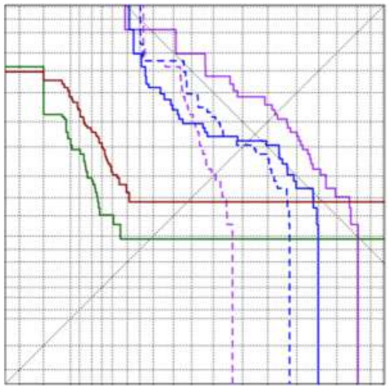
[redacted]

4 days of collect

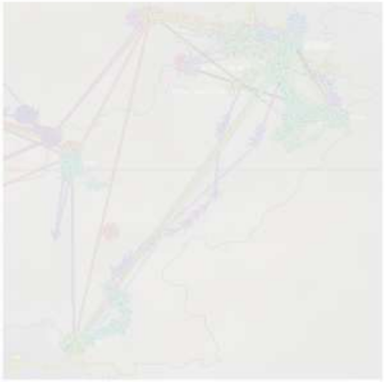
Now, we'll describe a cross validation experiment on the AQSL selectors that we were provided



Behavioral Feature Extraction



Cross Validation Experiment
on AQSL Couriers

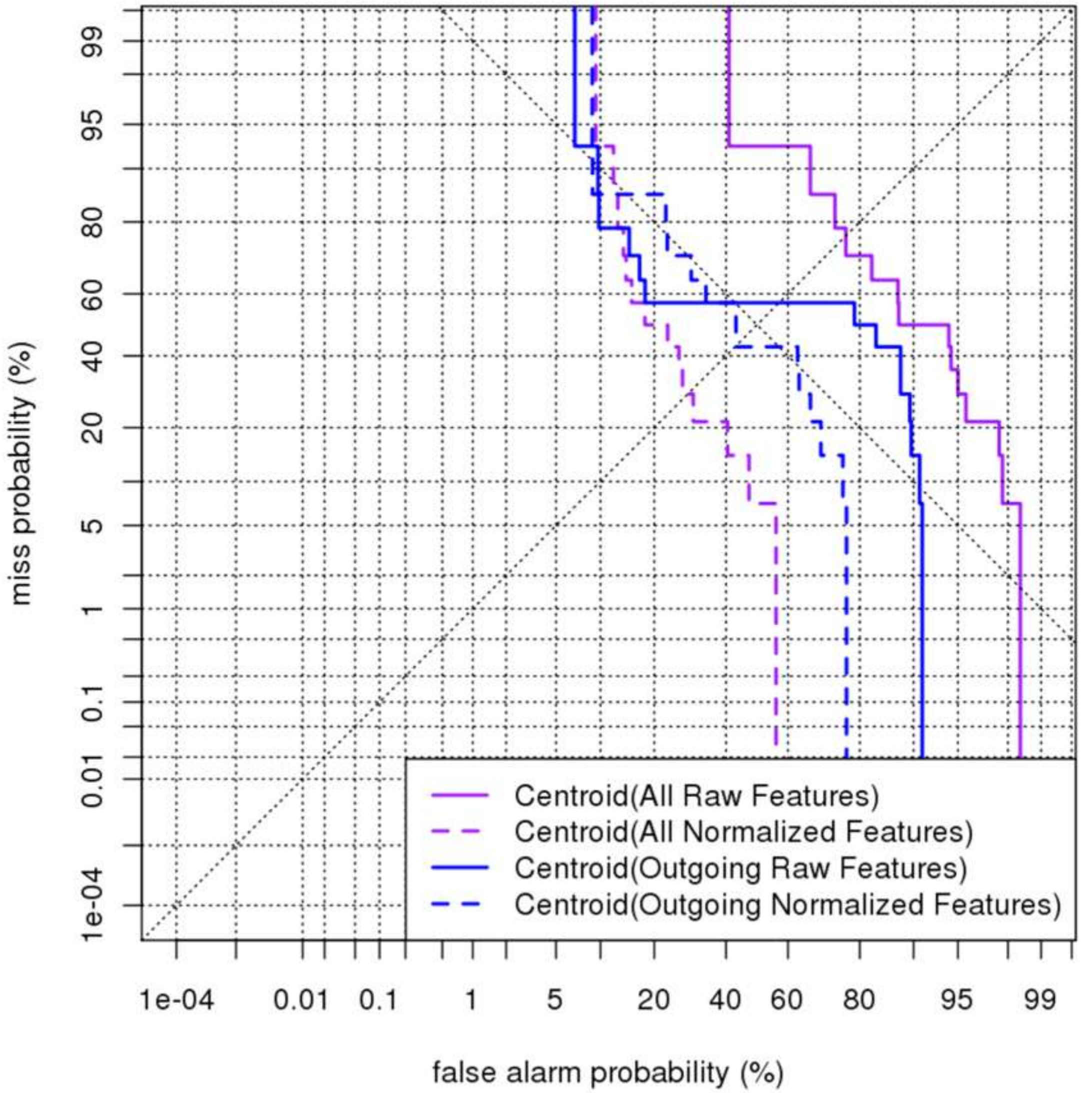


Preliminary SIGINT Findings

Our initial detector uses the centroid of the AQSL couriers to “find other selectors like these”

AQSL Cross-Validation Experiment

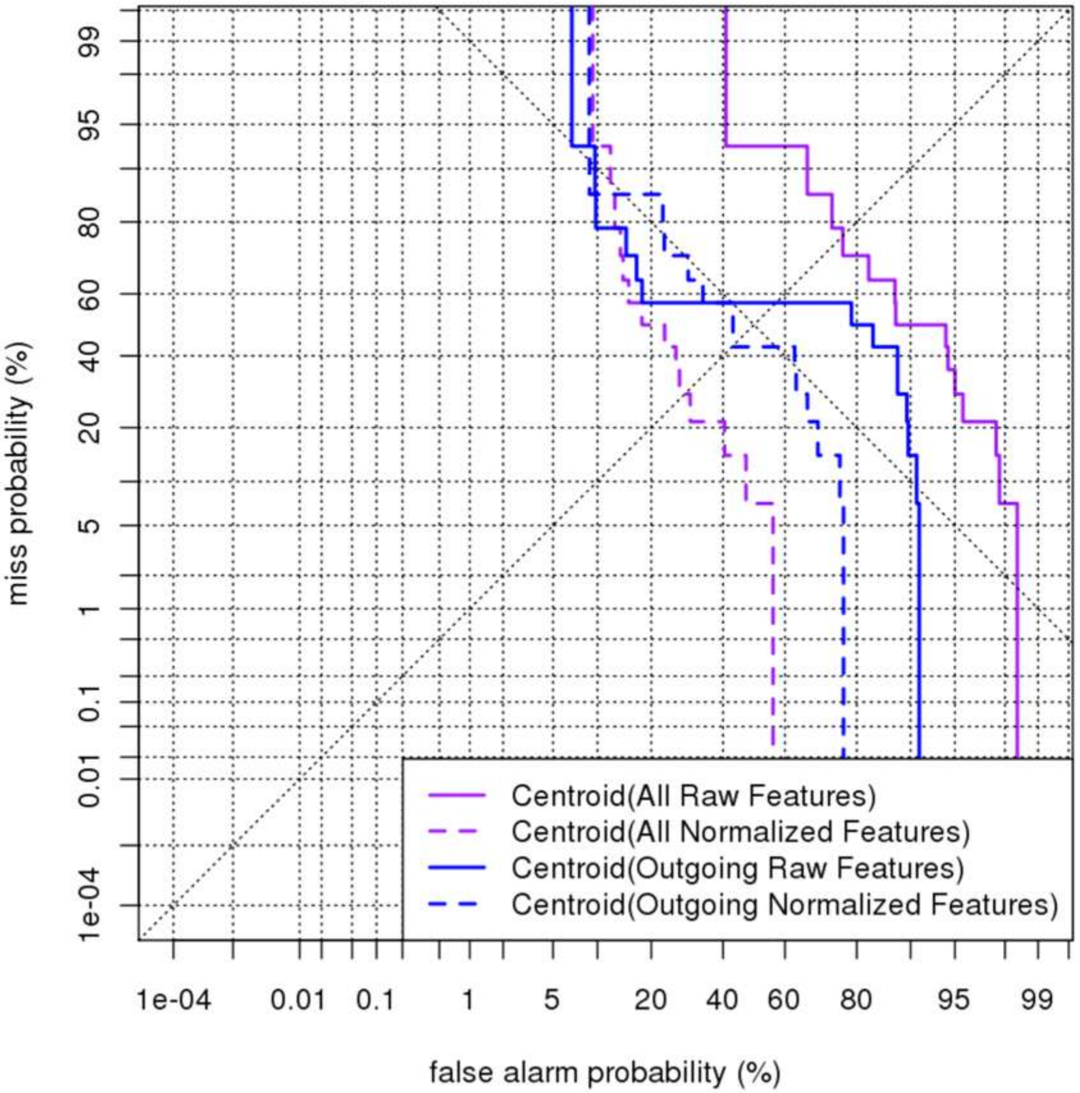
- 7 MSISDN/IMSI pairs
- Hold each pair out and score them when training the centroid on the rest
- Assume that random draws of Pakistani selectors are nontargets
- How well do we do?



Our initial detector uses the centroid of the AQSL couriers to "find other selectors like these"

AQSL Cross-Validation Experiment

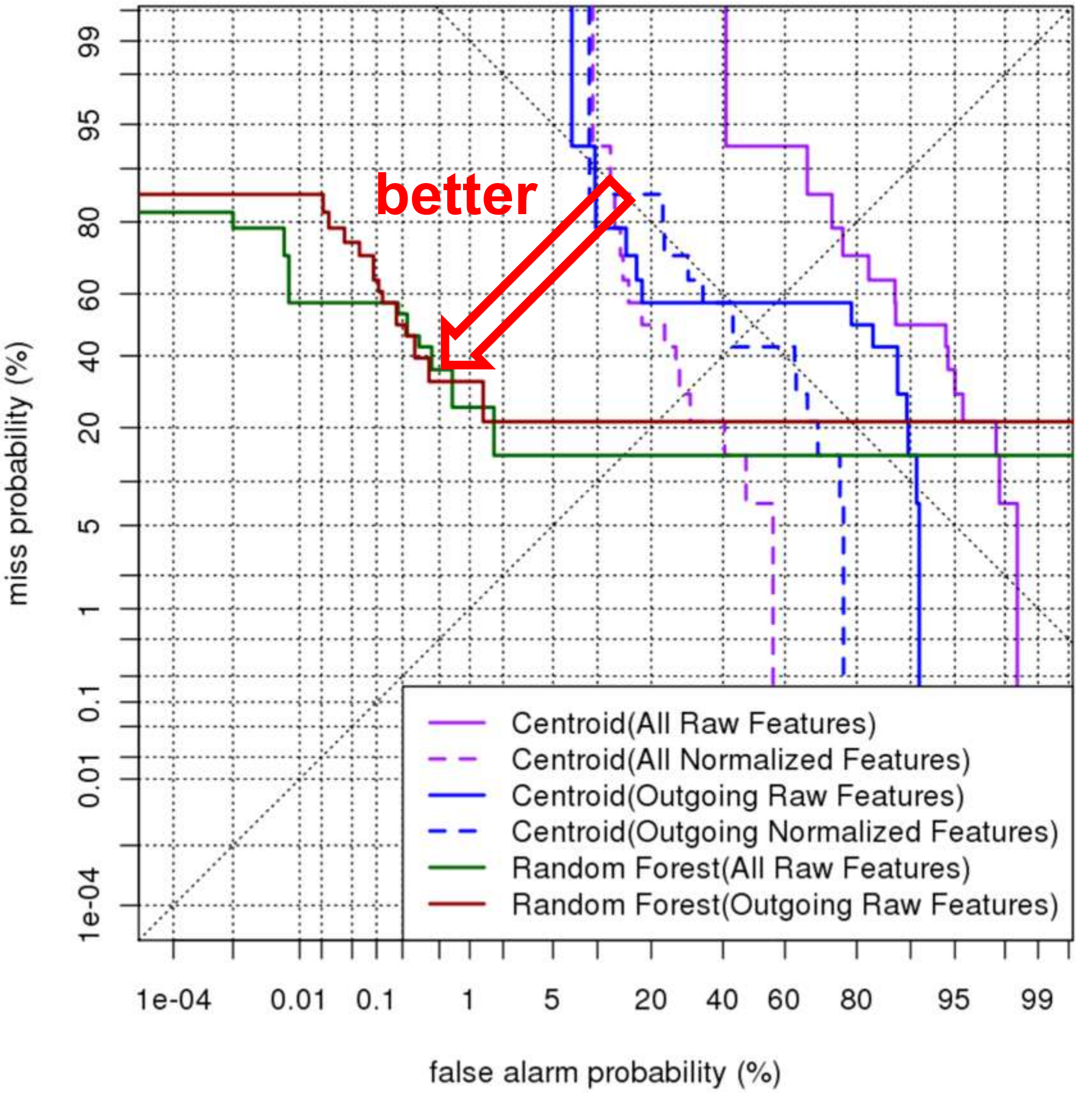
- Initial experiments showed EER in 10-20% range
- Here, performance is much worse against these nontargets:
  - Seen in Pakistan
  - Not seen outside of Af/Pak
  - Not FVEY selectors



Statistical algorithms are able to find the couriers at very low false alarm rates, if we're allowed to miss half of them

Random Forest Classifier

- 7 MSISDN/IMSI pairs
- Hold each pair out and then try to find them after learning how to distinguish remaining couriers fro n other Pakistanis
  (using 100k random selectors here)
- Assume that random draws of Pakistani selectors are nontargets
- 0.18% False Alarm Rate at 50% Miss Rate



We’ve been experimenting with several error metrics on both small and large test sets

Training Data	Classifier	Features	100k Test Selectors		55M Test Selectors	
			False Alarm Rate at 50% Miss Rate	Mean Reciprocal Rank	Tasked Selectors in Top 500	Tasked Selectors in Top 100
None	Random	None	50%	1/23k (simulated)	0.64 (active/Pak)	0.13 (active/Pak)
Known Couriers	Centroid	All	20%	1/18k		
		Outgoing	43%	1/27k		
	Random Forest		0.18%	1/9.9	5	1
+ Anchory Selectors						

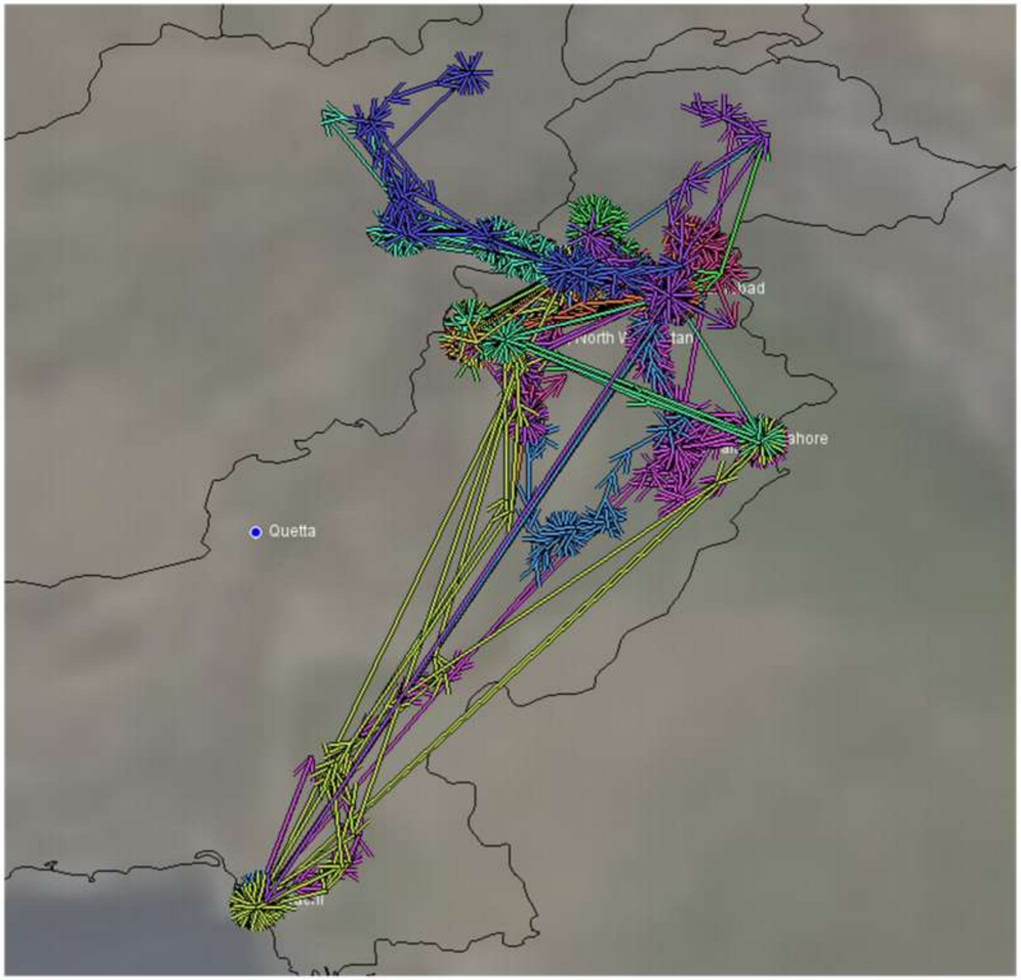
Random Forest:

- 0.18% false alarm rate at 50% miss rate
- 7x improvement over random performance when evaluating its tasked precision at 100

To get more training data we scraped selectors from S2I11 Anchory reports containing keyword “courier”

Anchory Selectors

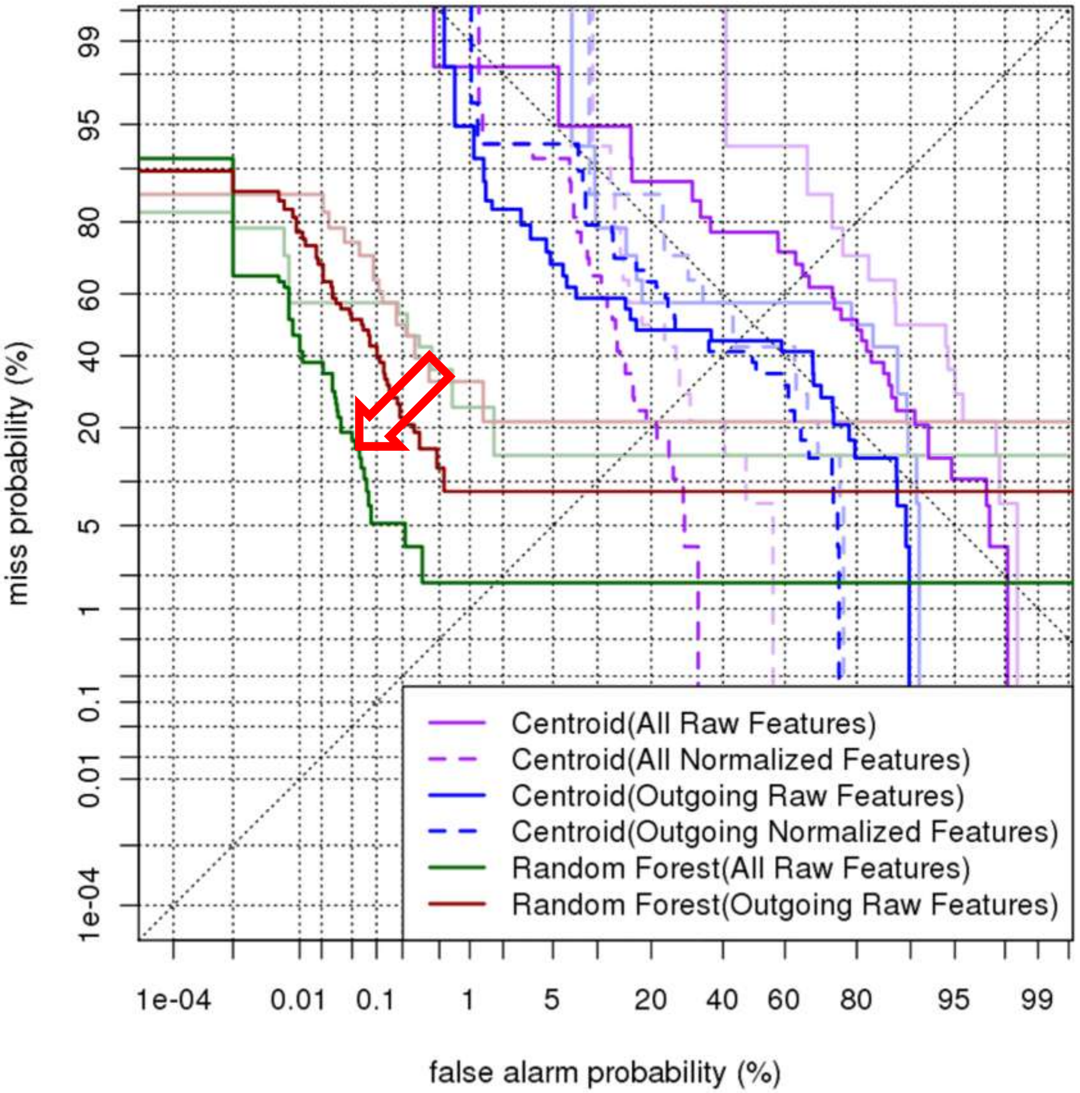
- Searched for reports containing “S2I11” AND “courier”
- Filtered out non-mobile numbers and kept selectors with “interesting” travel patterns seen in SmartTracker



Adding selectors from Anchory reports to the training data reduced the false alarm rates even further

Anchory Selectors

- Searched for reports containing “S2I11” AND “courier”
- Filtered out non-mobile numbers and kept selectors with “interesting” travel patterns seen in SmartTracker



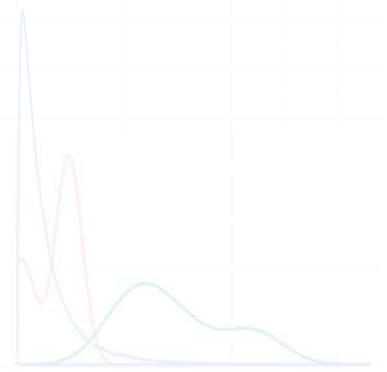
We’ve been experimenting with several error metrics on both small and large test sets

Training Data	Classifier	Features	100k Test Selectors		55M Test Selectors	
			False Alarm Rate at 50% Miss Rate	Mean Reciprocal Rank	Tasked Selectors in Top 500	Tasked Selectors in Top 100
None	Random	None	50%	1/23k (simulated)	0.64 (active/Pak)	0.13 (active/Pak)
Known Couriers	Centroid	All	20%	1/18k		
		Outgoing	43%	1/27k		
	Random Forest		0.18%	1/9.9	5	1
+ Anchory Selectors			0.008%	1/14	21	6

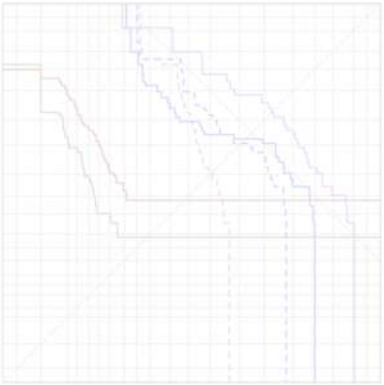
Random Forest trained on Known Couriers + Anchory Selectors:

- 0.008% false alarm rate at 50% miss rate
- 46x improvement over random performance when evaluating its tasked precision at 100

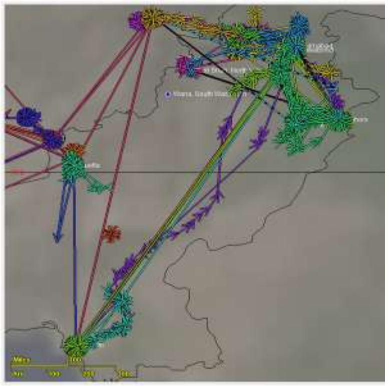
Now, we’ll investigate some findings after running these classifiers on +55M Pakistani selectors via MapReduce



Behavioral Feature Extraction

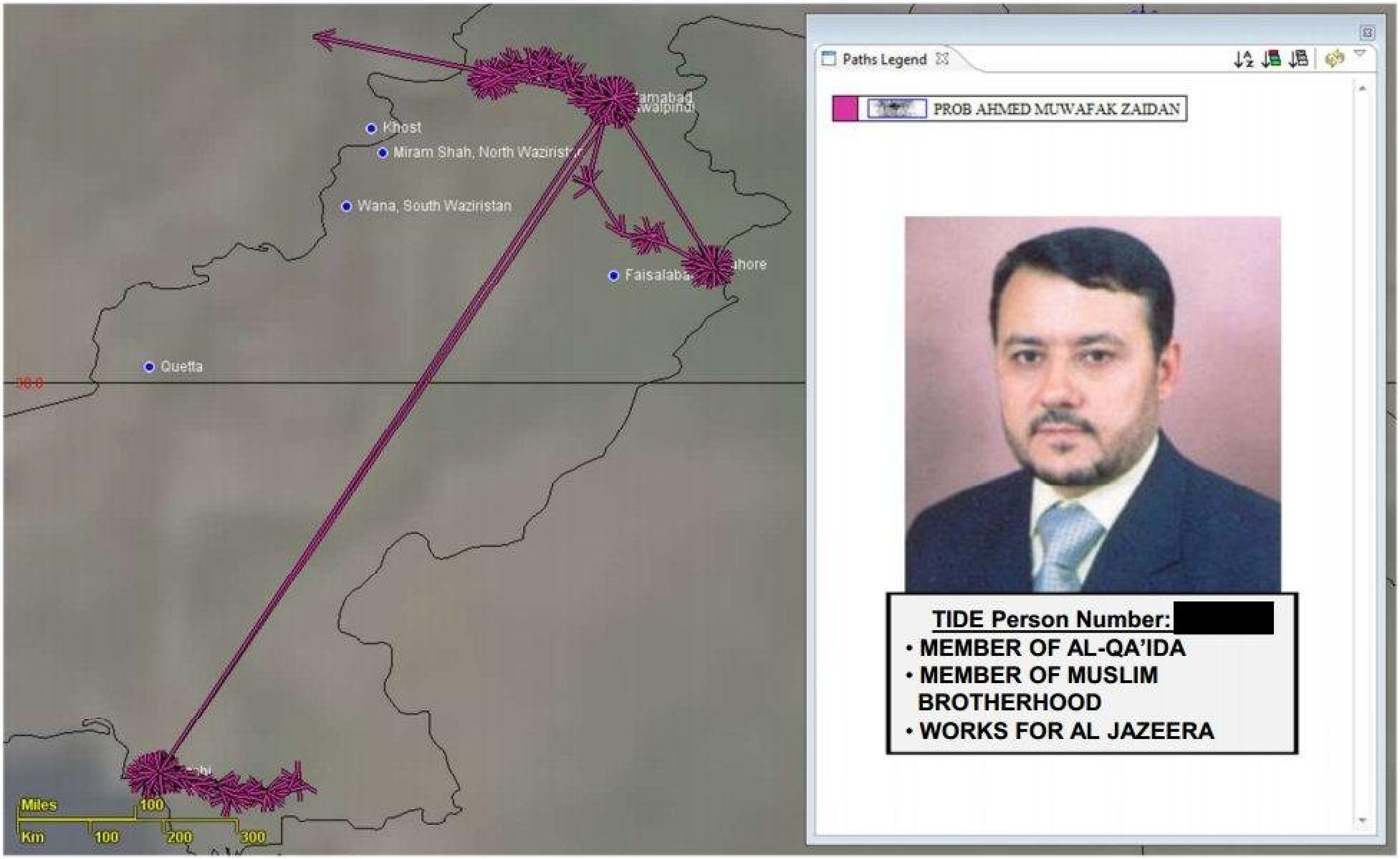


Cross Validation Experiment on AQSL Couriers



Preliminary SIGINT Findings

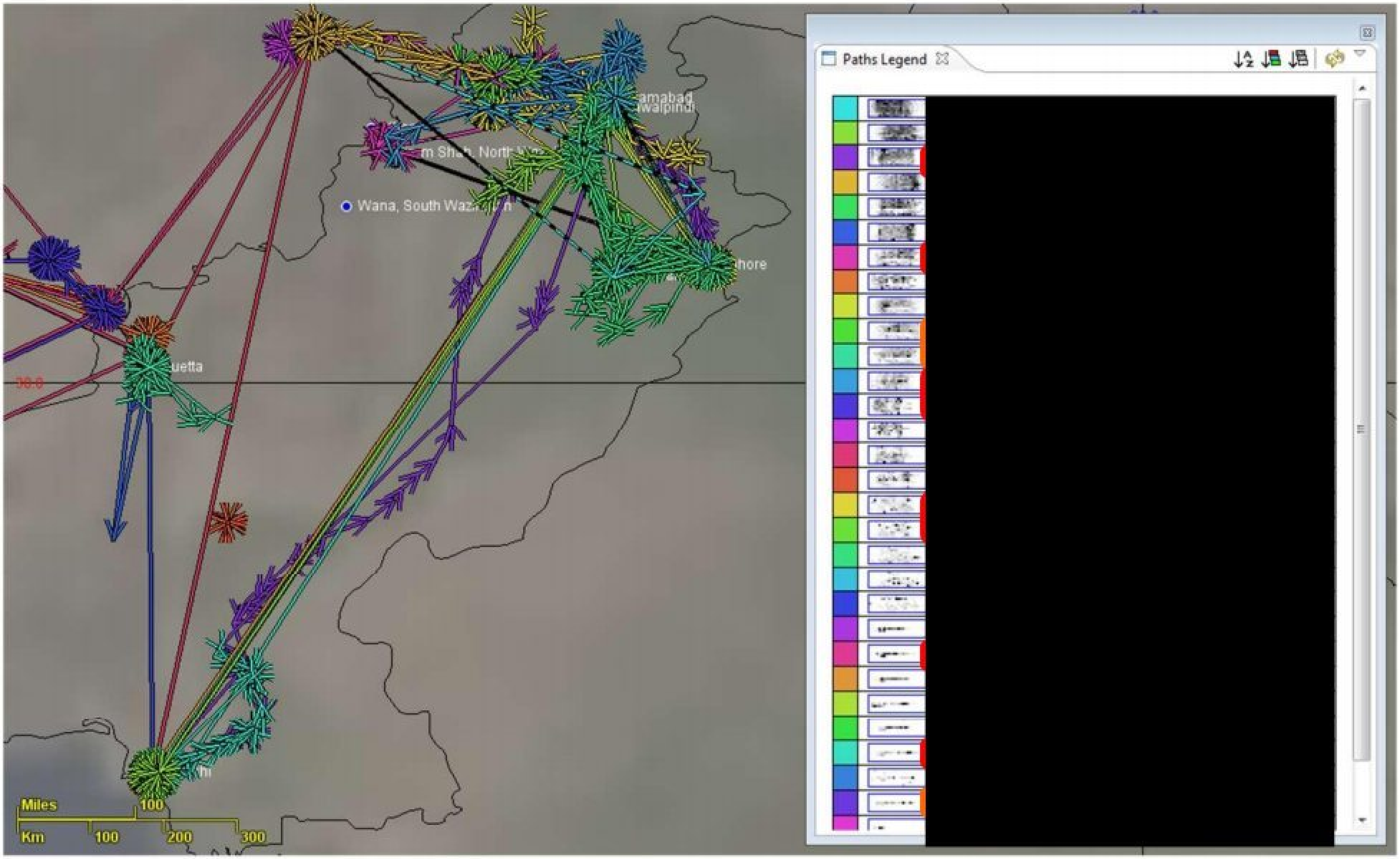
The highest scoring selector that traveled to Peshawar and Lahore is PROB AHMED ZAIDAN



TIDE Person Number: [redacted]

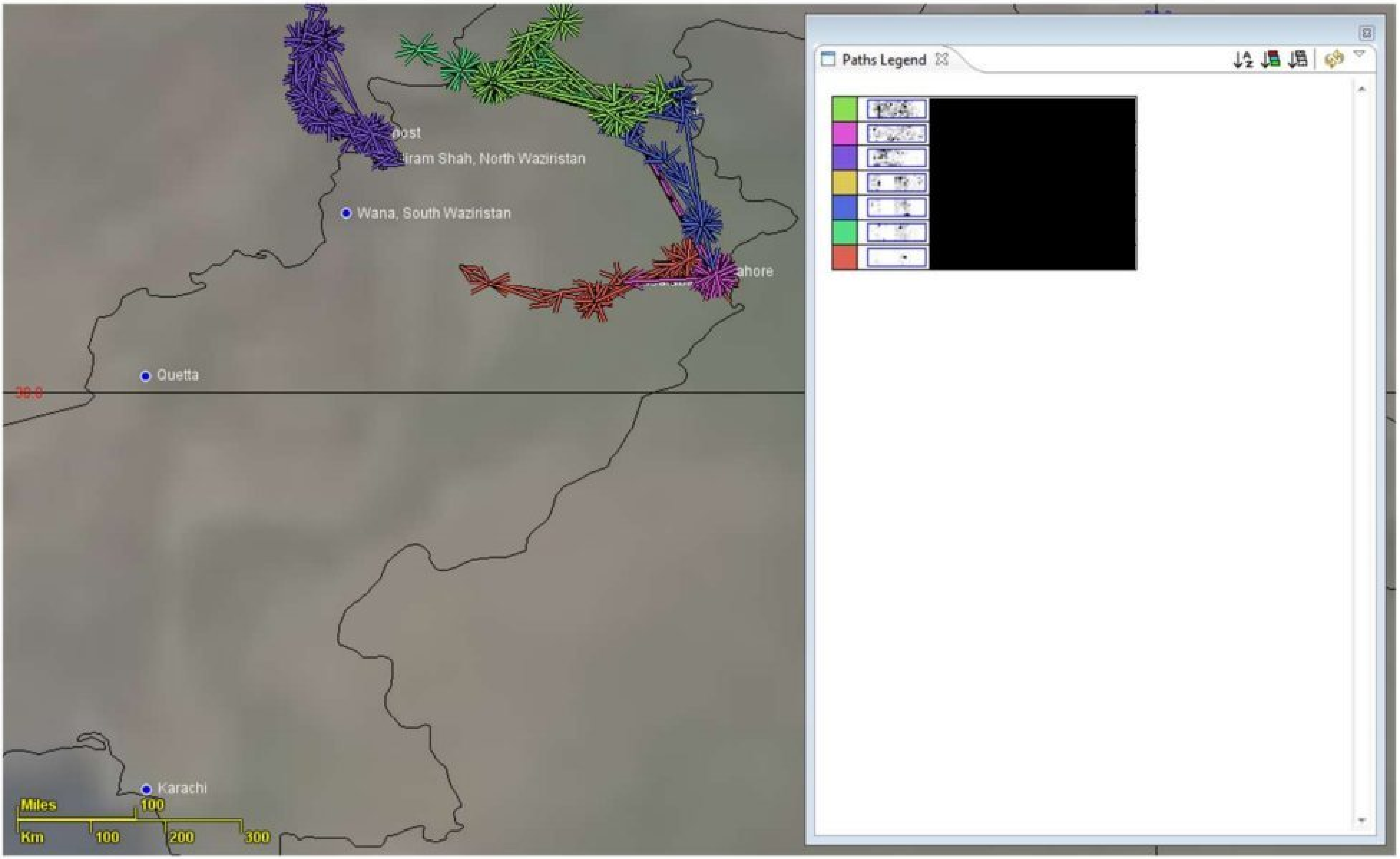
- MEMBER OF AL-QA'IDA
- MEMBER OF MUSLIM BROTHERHOOD
- WORKS FOR AL JAZEERA

In the top 500 scoring selectors, 21 are tasked leading us to believe that we're on the right track



[redacted]

We have also discovered many untasked selectors with interesting travel patterns

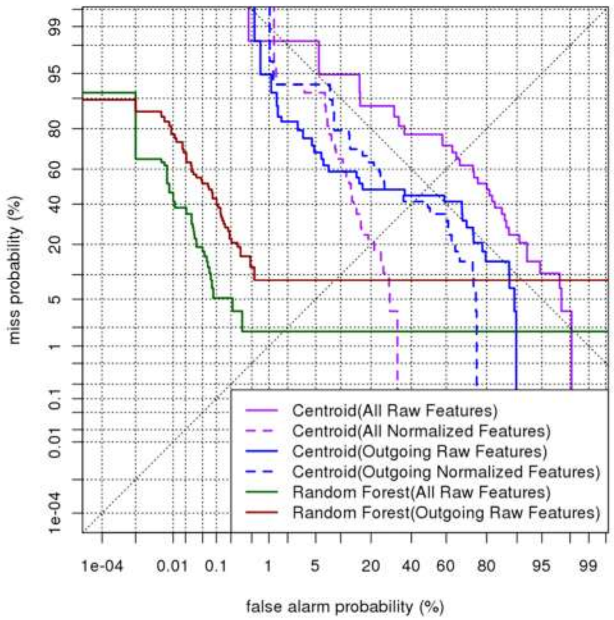


🟩	[redacted]	[redacted]
🟪	[redacted]	
🟪	[redacted]	
🟨	[redacted]	
🟦	[redacted]	
🟩	[redacted]	
🟥	[redacted]	

Preliminary results indicate that we’re on the right track, but much remains to be done

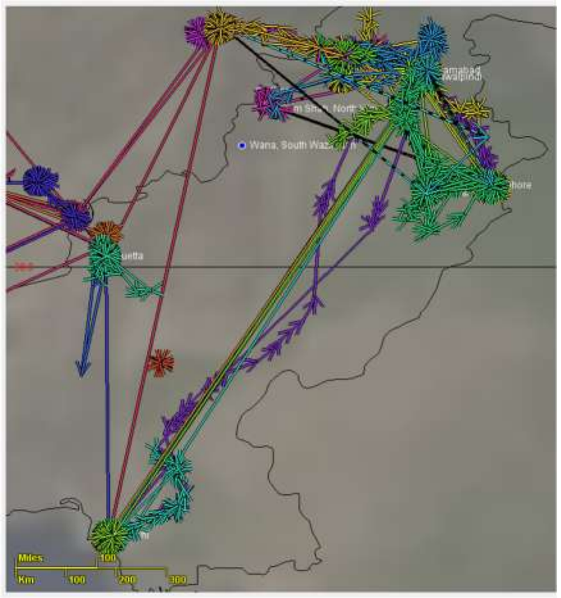
Cross Validation Experiment:

- Random Forest classifier operating at 0.18% false alarm rate at 50% miss
- Enhancing training data with Anchory selectors reduced that to 0.008%
- Mean Reciprocal Rank is ~1/10



Preliminary SIGINT Findings:

- Behavioral features helped discover similar selectors with “courier-like” travel patterns
- High number of tasked selectors at the top is hopefully indicative of the detector performing well “in the wild”


Wana, South Waziristan
uetta
hore
Miles
100
Km
100
200
300