Summary of DNR and DNI Co-Travel Analytics

S2I5

1 October 2012

Contact information:▮

Table of Contents

Introduction

(S//REL TO USA, FVEY) This short-term study overviews and documents key elements of the co-traveler analytics both under development and operational at NSA. Each section includes a brief description of the analytic, its status, source data, and caveats.

(S//REL TO USA, FVEY) While each analytic was designed to operate on a particular type of data or a particular data format, many can likely be scaled to operate on other data sources. For instance, analytics designed for DNR GCID or VLR data might also apply to DNI Geolocation data.

(S//REL TO USA, FVEY) The process of documenting these analytics raised a series of important issues that not only distinguish the analytics from each other, but more importantly, shape the landscape that we must consider in moving forward to meet the analytic needs at NSA. Some of these issues are discussed in the next section.

Issues and Questions

Should a co-travel analytic consider where a GCID or VLR is physically located?

- o   Many GSM analytics use GCID information to identify co-travelers. If two selectors are seen at the same GCID around the same time, they are considered co-travel candidates. The analytic does not need to know where the GCID is physically located. However, if the individuals are using different network providers (e.g., T-Mobile and Verizon), they may be physically standing next to each other as their mobiles register with different cell towers. Co-travel analytics that do not consider the physical geo-locations of the towers will not discover individuals that are co-traveling on different networks.
- o   Analytics that make use of point data (e.g., Thuraya) necessarily need to consider geolocational data in order to determine distance from one point to another.

Should incidental co-travelers be considered?

- o   There is a difference between incidental co-travel due to collective movement (individuals with similar travel behaviors but no other similarities) and functional group-based co-travel among individuals with behaviorally relevant relationships. CTCOP makes this definition explicit, but warns that we might not want to exclude seemingly incidental co-travelers simply because we are unaware of their relationship.
- o   Other factors, such as contact chains and target COMSEC behaviors (frequent power-down, handset swapping, SMS behavior), might assist in determining whether co-travelers are associated through their travel behaviors alone or through behaviorally relevant relationships.

Should geography play a role in co-travel?

- o Because it is difficult to know where a GSM target is located within a GCID or VLR, many of the GSM co-travel analytics use the mathematical central point in the VLR or GCID as a reference point. We could postulate that traveling targets will be located along roads, train tracks, or footpaths where network service exists. This type of geographical information could theoretically be used to inform a co-traveler analytic in identifying candidates (especially those that are traveling via the same means of transportation). Geographical information might also be used to "fill in the gaps" when data is missing between locations that a target visited.
- o Analytics in this study that make use of such geographical information include DSD's Co-travel analytic and the Geospatial Analysis Tradecraft Center's (GATC's) Opportunity Volume analytic.

Should device and collection sampling play a role in determining co-travelers?

- o We may collect hundreds of events from one target's mobile phone while collecting only a few events from his co-traveler's mobile phone. The number of events collected may be due to collection bias, differences in network service, and/or target COMSEC behavior. Analytics should take these considerations into account when attempting to identify co-travelers.

Should co-travelers seen in different source databases be considered?

- o Depending on a target's preferred communication behaviors, some co-travelers may be seen largely in DNR GSM data, and other co-travelers may be seen largely in DNI data. We may be able to construct a more complete picture of a target's locations over time if we combine DNR and DNI data sources. It might be worth considering the degree to which considering multiple data sources will significantly increase the number of false positives.
- o Databases that do not contain geolocation information might also be considered. For instance, air travelers on the same reservation number are probably co-traveling on the same flight. Users sharing a MAC address are probably co-located using the same device even though we may not know where that device is located. Consistent observations of devices within the same LAIC may provide evidence of co-location, even if the LAIC's physical service area is unknown. Finally, similarities between IP addresses may indicate proximity on the same LAN, even if the physical location of the LAN nodes is unknown.
- o The one analytic in this study that attempts to combine multiple sources of information to build a more holistic picture of a target's travel pattern is the TAC/Café/TMAC Co-travel analytic.

Can co-travel be considered a series of meetings?

- o We attempted to limit this study to targets co-traveling through two or more locations within an analyst-specified time and space window. If those locations are defined, however, we might consider co-travel as a series of "meetings" at known locations. Analytics that detect co-location may be different in nature from those that detect co-travel. The specific analytic need will define which of these approaches is more appropriate and efficient.

o In this study, examples of meeting analytics that detect instances of co-location include the GATC Opportunity Volume Analytic and the ████ Meet&Greet Spatial Chaining Analytic.

Analytics

CHALKFUN

████████████████████

Background

(TS//SI/REL TO USA, FVEY) Chalkfun's Co-Travel analytic computes the date, time, and network location of a mobile phone over a given time period, and then looks for other mobile phones that were seen in the same network locations around a one hour time window. When a selector was seen at the same location (e.g., VLR) during the time window, the algorithm will reduce processing time by choosing a few events to match over the time period. Chalkfun is SPCMA enabled[1].

(S//SI/REL TO USA, FVEY) Note: As of 6 September 2012, the events that are chosen depend on the "sampling method" chosen by the analyst (most active, most per day, first/last/most, or first/last/spread). The "sampling rate" specifies how many events are chosen to match. As Chalkfun moves to the cloud, this option will be discontinued.

(TS//SI/REL TO USA, FVEY) The cloud-based version of Chalkfun (see R6 SORTINGLEAD Co-traveler Analytic section), which may be released as early as September 2012, will have a number of additional features and options:

- The system will run one query (rather than separate queries) for all of the IMSIs, MSISDNs, VLRs, and GCIDs that an analyst enters (as if the selectors and areas of interest were joined with an "OR"). The system currently runs separate queries for each, returning separate sets of results for each combination of selector and areas of interest. The cloud-based version will also enable the user to set the size of the time window that the analytic considers, rather than defaulting to one hour (as described above).
- The user will be able to choose the countries or locations of interest. Blacklist and whitelist features will enable the user to instruct the system to ignore activity within a region, or restrict analysis to specified regions of interest (e.g., ignore activity in ██████ or use only activity from ██████)
- In considering potential co-travelers, the analyst will have the option to ignore activity in which the target is in his home country

---

[1] (S//SI//REL) SPCMA enables the analytic to chain "from," "through," or "to" communications metadata fields without regard to the nationality or location of the communicants, and users may view those same communications metadata fields in an unmasked form.

- The analyst will be able to filter in or out potential co-travelers with specified prefixes (for instance, return only ▮▮▮▮ mobiles, remove all ▮▮▮▮ mobiles, them, or include only mobiles that are from the same country as the target).

Status and Summary

Status	Source Data	Caveats
- Operational; Available at analysts desktops - Cloud version could be available as early as September 2012.	- All FASCIA data containing VLR and GCID information	- Current version is not cloud-based and can have long processing times, however cloud-based solution is imminent. - Analytic will only return co-travelers on the same provider network

DSD Co-Travel Analytic

▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮▮

Background

(S//SI/REL TO USA, FVEY) The DSD Co-Travel analytic predicts target locations and co-travelers by calculating time-based travel trajectories. Probable travel routes are calculated using observed locations and determining the most likely paths and travel times similar to that used in turn-by-turn navigation systems. These target travel paths are represented as a series of LAT/LONG waypoints or line segments along the probable travel routes, such as roads. The travel paths are divided into segments (e.g. 20 to 50km along the road). The analytic predicts the approximate time that the target would theoretically arrive at each segment waypoint based on projected travel times between known locations. Then, within the travel window, the analytic discovers candidate co-travellers that intersect locations along the buffered travel path. The next step in the analytic is performed using interactive Renoir analysis of a two mode graph representing the route segments and selectors observed on these route segments within the time windows. Once the data is clean and candidate co-travellers are identified detailed analysis can be done in Renoir or other tools such as GeoTime incorporating other supporting data such as communications events and content.

(S//SI/REL TO USA, FVEY) The analytic currently runs on a Netezza-based architecture, called Hectic Snare, that rapidly executes MySQL-based QFDs. This architecture enables interactive exploratory analysis and rapid pattern matching. The analytic is distributable and could be implemented in Hadoop/MapReduce or Accumulo.

(S//SI/REL TO USA, FVEY) This analytic was tested using an ▮▮▮▮▮▮ terrorist case study. The case study used approximately 80,000 base stations locations and 16 billion mobiles location records for CDRs (Call detail records) and infrastructure collect from DRT and Juggernaut systems. This case study showed that more candidate co-travellers were discovered by analyzing the travel paths than by considering common meeting locations alone.

Status and Summary

Status	Source Data	Caveats
Analytic implemented and tested at DSD.	- Mobile CDRs and residing in Netezza-based architecture.	- Requires Netezza (current implementation) - Requires Renoir

Future Work

(S//SI/REL TO USA, FVEY) DSD would like to integrate key meeting locations into this analytic, such as safehouses. Plans are also underway to identify targets based on COMSEC behaviors such as identifying mobiles that are turned off right before convergence between two travel paths occurs.

Geospatial Analysis Tradecraft Center (GATC) Opportunity Volume Analytic

█████████████████████████████

Background

(TS//SI/REL TO USA, FVEY) The opportunity volume analytic determines whether two entities (e.g. devices) could have been co-located by considering the possibility of their travel paths intersecting. The opportunity volume analytic requires pairs of event locations and times for each entity, and computes the possible locations and times in which the two entities could have been co-located. It does this by computing possible travel route surfaces for each entity between the specified events, using a travel cost surface computed from terrain, land cover, and road network data. These possible travel route surfaces include the temporal dimension (that is, the period of time in which the entity could have been at the given location); the intersection between these multidimensional surfaces represents the places and times during which the entities could have been co-located. The analytic was developed using GPS point event data, but the analytic actually uses a 1-km grid for the spatial resolution and a 15-minute period for the temporal resolution, so it can be applied to any data that can be expressed in these terms.

Status and Summary

Status	Source Data	Caveats
Prototype service implemented on NGANet. Not yet ported to NSANet.	- Geohashes of GPS point event data.	- Requires event locations and times for every selector. - Designed for 1 km grid-based locations and 15 minute time intervals. - Co-travel capability would require analyst to define a series of meetings at specified locations.

Future Work

(TS//SI/REL TO USA, FVEY) The purpose of this service is to determine whether two entities could have been co-located given observed event locations for those entities. To detect co-travel, the analyst would need to define a series of meeting locations and times. The opportunity volume analytic could also provide a mechanism for vetting co-travel analytics by testing for possible co-location events along co-travel routes.

███████████████ TMI Co-Traveler Analytic
███████████████████████████

Background

(TS//SI/REL TO USA, FVEY) The ████████████████ Track Mutual Information (TMI) cloud analytic was developed as a study under their graph analytics, alerting, and target development program. The analytic is oriented to work on 7 to 30 days worth of regional collection. It has been tested on RT-RG data from the ██████ region. Instead of using GCID information as co-travel reference points, the analytic works cross-network by computing target "closeness" based on the GCID Lat/Long GEO information and time. The Lat/Long information is obtained from RT-RG.

(TS//SI/REL TO USA, FVEY) The analytic starts by computing event sequences of LAT, LONG, and time for each selector. These are called "tracks". It then computes a value that measures how far the selector has traveled in general. If the selector has not traveled outside a 20 to 50 km radius, the selector is not considered. Each eligible selector's tracks are pairwise-compared to the others and a measure of similarity in time and space is computed.

Status and Summary

Status	Source Data	Caveats
Initial development completed. In testing phase, not yet operational	- Sortinglead summaries of FASCIA data on GM-PLACE and GM-████ - RT-RG regional GSM collection ████████	- Analytic only considers tasked selectors as seeds. - Analytic does not consider targets that do not travel outside a 20 to 50 km radius. - Track dataset must be repopulated for each data update

Future Work

(TS//SI/REL TO USA, FVEY) █ would like to reduce processing by creating an index containing selectors whose tracks are near each other in space. To achieve this, future work may make use of a GEOAddress hashing algorithm that uses LAT/LONG information to group cell towers into clusters that are in the same region. This hash considers latitude and longitude only, and is agnostic to the targets' service provider. It may be possible to also compare target tracks quickly by comparing these GeoAddresses.

████ Co-Traveler Analytics

████████████████████████████████

Background

(TS//SI/REL TO USA, FVEY)████ has developed two co-travel analytics: Fast Follower (FF) and Meet&Greet Spatial Chaining (MGSC). The FF analytic was initially designed to detect individuals who are following station personnel. Detailed non-SIGINT path data is collected consensually on the station personnel, and this reference path data provides the seeds for this analytic, which attempts to discover mobile GEO data indicating individuals that may be following the station personnel. The MGSC analytic is designed to detect meetings between high-value individuals and other entities.

(TS//SI/REL TO USA, FVEY) The FF analytic begins by considering non-SIGINT reference paths for station personnel based on detailed knowledge of the entity's location. Candidate followers are determined by identifying other individuals that have traversed some number of consecutive points (determined by the analyst) that match the reference path in space and time. The analyst also sets a parameter to specify the minimum distance that must be covered along a candidate path.

(S//SI/REL TO USA, FVEY) The MGSC analytic is designed for ELKPRINTS data from smartphones. This analytic identifies sequences of consecutive location points close in time and combines them into a single data point. A maximum velocity movement parameter is applied to create a time window around each point representing the approximate time at which the individual was located there (as opposed to traveling to or from that location). Finally, co-travelers are identified by discovering pairs of selectors that meet the duration and distance thresholds set by the analyst as input parameters. Spatial chaining software aggregates and presents the meeting data, including the locations, times, and scoring metrics to the analyst.

Status and Summary

Status	Source Data	Caveats
The MGSC analytics has been tested on real ELKPRINTS data, but results have not been validated by operational analysts. **The FF analytic has been tested on made-up data.**	- Smartphone data from ELKPRINTS - Reference-path data (FF) - List of selectors (MGSC)	- Analytic designed for precise geolocation data (e.g., from smartphones) - MGSC analytic would require the analyst to define a series of meetings

PACT NGA-NSA GATC Analytic

████████████████████████████████

████████████████████████████

Background

(TS//SI/REL TO USA, FVEY) The PACT analytic is a joint NSA-NGA effort to identify co-traveling Thuraya handsets. The effort was motivated by an increase in Thuraya phone usage by ████████████ ████████████████. SIGINT Geospatial Analysts were able to characterize the travel

behaviors of the targeted Thuraya handsets and identifying other handsets with similar patterns. The targeted handsets were observed traveling between known ████ government and military installations; therefore, handsets with similar travel behaviors were inferred to be ████ government forces.

(TS//SI/REL TO USA, FVEY) The first step of PACT is to identify a set of waypoints for each target handset. Waypoints are generated from sequences of events that cluster together in space and time. The second step is to identify which pairs of handsets contain similar waypoint clusters. Pairs are scored based on the number of waypoint clusters that match. This analytic also considers the total possible number of waypoint clusters for each selector, so that the total number of communication events per selector is taken into consideration. This process is intended to reduce the possibility of producing results that include incidental co-travel. The third step in this analytic identifies persistent patterns by examining the time periods over which co-location occurs for each co-travel candidate pair.

Status and Summary

Status	Source Data	Caveats
Tested on VOICESAIL data from CULTWEAVE. Patterns stored in QFD. **In process of transitioning PACT to NSA/S2.**	- Thuraya data from CULTWEAVE (~500 M waypoints in CULTWEAVE)	- Analytic designed for Thuraya or other point data

Future Work

Future work could involve applying this analytic to other types of QFD datasets such as Inmarsat and GSM data. The team is also interested in building on this analytic to enable discovery of asynchronous co-traveling relationships.

R6 SORTINGLEAD Co-Traveler Analytic

████████████████████████████████████

Background

(S//REL TO USA, FVEY) R6 has been partnering with Chalkfun to upgrade the Chalkfun co-traveler analytic to a cloud-based analytic that will run on Cloud 14 (to eventually be migrated to MDR-2).

(TS//SI/REL TO USA, FVEY) The R6 co-traveler analytic accepts a selector and timeframe as input, and then derives an itinerary for the selector that includes the CELL IDs and/or VLRs (depending on what is available). The itinerary is based on a series of waypoints generated from the location information that is available in FASCIA-PCS. Then, the analytic searches for other selectors that were "near" these waypoints in space and time. Time windows are configurable and can be adjusted by the user. Each candidate is scored and then prioritized based on the scores.

(TS//SI/REL TO USA, FVEY) The R6 co-traveler analytic operates on Sortinglead Event Summaries and a GEO Index. The Sortinglead Event Summaries provide rapid access to FASCIA PCS events by summarizing

and enriching key elements of selector behavior. The Sortinglead Event Summaries benefit this analytic because they can provide enriched location information about selectors that is not present in the raw metadata. The GEO Index contains a mapping between the locations (GCIDs or VLRs) visited by a selector and the time (day/minute) that the visit(s) occurred. Information from command and control networks that track IED attacks is also used to enrich the GEO Index.

(TS//SI/REL TO USA, FVEY) The results that can be returned from this type of analytic can potentially be enormous. Each candidate will have some level of time and space overlap with the seed. Prioritization occurs by assessing the quality of the overlap in terms of time and space closeness. The analyst may choose to triage any number of potential candidates (e.g. top 10 or top 100 candidates, or candidates that surpass a given threshold).

Status and Summary

Status	Source Data	Caveats
- In testing phase to be replacement back-end for the current production CHALKFUN co-traveler tool **- Cloud-based (MapReduce) implementation under development to handle larger numbers of queries simultaneously**	- FASCIA PCS Sortinglead Summaries - CHALKFUN enrichment (VLR country mapping)	- Analytic cannot recover cross-network co-travelers - Analytic will not be effective against stationary (non-traveling) targets - Processing is memory intensive - Analytic is sensitive to large cells, VLRs, and dense areas - Not directly applicable to sat phones with LAT/LONG information - Results can be very sensitive to timeframe chosen as input. For instance, analytic will not be effective for large queries across multiple countries and large time frames (e.g., anywhere in ██████ over the past year and then anywhere in ██████).

Future Work

(TS//SI/REL TO USA, FVEY) Because the R6 co-traveler analytic depends on GCID and VLR locations as meeting points or waypoints, it will not return selectors that co-travel on different provider networks. (For instance, it could not return a Verizon selector co-traveling with a T-Mobile selector.) The R6 team is working on experiments that might "alias" seed selectors to nearby selectors on other networks to get around this problem, but this poses challenges. The RT-RG analytic (discussed later in this paper) uses relative velocities to deal with the cross-network challenge, but this approach requires pre-computing travel behavior for all pairs of selectors, which can be computationally expensive.

RT-RG Sidekicks

███████████████████████████

Background

(TS//SI/REL TO USA, FVEY) The RT-RG Sidekick Cloud-Based Co-traveler analytic compares average travel velocity between pairs of selectors to infer whether or not could co-travel would practically be possible. The velocity factor is intended to reduce the number of false positives when considering travel among urban areas by filtering out pairs of selectors that were seen at the same series of CELL IDs or VLRs over time, but could not have been traveling together because the location data timestamps presuppose an unreasonable velocity. This may happen because one or both of the selectors in the pair may have been located at the edges of the network coverage during one or more of their travel midpoints.

(TS//SI/REL TO USA, FVEY) The analytic first computes "movement summaries" of all available tasked selectors. The movement summaries contain a list of locations that a target visited during the timeframe of interest, given by the analyst. Locations are defined by CELL IDs (for GSM) or GEO-Hashes (for ████ any other selectors with Lat/Long). Then, the system discovers pairs of targets that could be traveling together by comparing their sequences of physical locations and factoring out pairs that could not have reasonably arrived at the meeting waypoints within 10 minutes of each other.

(TS//SI/REL TO USA, FVEY) One of the main benefits of the RT-RG Sidekicks analytic is that it is not constrained by provider network. Because it considers physical (LAT/LONG) locations and travel velocities, it can provide co-traveler results that include selectors on different provider networks.

Status and Summary

Status	Source Data	Caveats
- QFD available at RT-RG analyst desktop. **RT-RG Tools: Goldminer, CHET, GEOT**	- Sortinglead Event Summaries ████████████████ ████ ████ of Fascia PCS) - Currently running on RT-RG ████████ - Could possibly scale to FASCIA event summaries	- Requires accurate tower geo data (location and date) - Requires pre-computing all selectors against all selectors, which can be expensive - Current output includes only tasked selectors - Analytic is not designed for stationary targets.

Future Work

(TS//SI/REL TO USA, FVEY) Currently, the system is integrated with RT-RG, operating on ████████ GSM data. It may scale to a larger data source; however, it is designed to precompute sidekicks for each possible pair or tasked selectors.

(TS//SI/REL TO USA, FVEY) This analytic could also be applied to DNI location data.

Scalable Analytics Tradecraft Center (SATC) Geospatial Lifelines Co-Travel QFD

███

Background

(TS//SI/REL TO USA, FVEY) The geospatial lifelines QFD applies the concept of "dwell times" to identify DNR co-travelers. Dwell times describe the time period spent at the beginning or ending destination. A location is considered a beginning or ending location if the dwell time at that location is greater than 2 hours.

(TS//SI/REL TO USA, FVEY) This QFD first generates geohashes using GSM event data, and then calculates transition lines indicating that a device traveled from one geohash to another. The result is a graph in which the geohashes represent nodes and the transitions represent links or edges. Clustering algorithms are applied to the graphs to determine locations and selectors of interest.

(TS//SI/REL TO USA, FVEY) The geospatial lifelines represent the beginning and ending locations, as defined by their dwell times, and all other intermediate observations. The likeliness of co-travel along paths between starting and destination points is based on the following measurements: net distance, time of transition (mins), speed (kph), Azimuth, and number of travel segments.

Status and Summary

Status	Source Data	Caveats
Analytic tested on 90 days of GSM event data from ██████ ████████ **Code is available through SATC, but analytic is no longer under development.**	- Geohashes of GSM event data retrieved from FASCIA.	- Analytic designed for GSM data, but could be applied to other types of data - Oriented to targets that remain in one location for at least 2 hours - Requires Geocoded source data for generating Geohashes

Future Work

(S//REL TO USA, FVEY) The code for this QFD is available through SATC, but the analytic is no longer under development. Ideas for future work before the project ended included adding acceleration and sinuosity to the computation.

SSG Common IMSIs Analytic

███████████████████████████

Background

(S//SI//REL) The Common IMSIs Analytic is a model in SEDB JEMA finds SIM card activity seen on cell tower panels in multiple areas (e.g.- border crossings commonly used by traffickers). It makes use of the Tower QFD.

(S//SI//REL) Analyst inputs areas of interest and time range. The analytic returns an excel file with a list of IMSIs seen in those areas at that time. It is enriched with OCTAVE tasking information. Limitations are that tower locations in OCTSKYWARD can be imprecise. Also, the SEDB Tower QFD summarizes IMSIs by LAIC by day. Summaries by MSISDN or IMEI are not available.

Status and Summary

Status	Source Data	Caveats
Available in JEMA.	-OCTAVE and FASCIA	- Cell tower locations in OCTSKYWARD can be imprecise. - The SEDB Tower QFD summarizes IMSIs by LAIC by day. - Summaries by MSISDN or IMEI are not available.

Additional Information
https://wiki.nsa.ic.gov/wiki/Analytics_Taxonomy

https://wiki.nsa.ic.gov/wiki/DNR_Travel_Pattern

Target Analysis Center (TAC)/Café/ Travel and Mobility Analysis Center (TMAC) DNI Co-Travel Analytic
Café Spin 1 (October 2011 – January 2012)

██

███████████████████████████

Background
(TS//SI/REL TO USA, FVEY) The Café project involved TMAC, SSG, T1212, and S2I5 working in concert to develop both DNI and DNR cloud-based travel analytics. The absence of a cloud-based solution that could run over bulk data motivated this initiative. The Café objective was to steer cloud travel analytics toward operational use and ultimately merge the DNI and DNR analytics in a unified co-travel analytic. These analytics are currently still under development; however, they are available to the development community on GM-PLACE.

(TS//SI/REL TO USA, FVEY) This analytic uses IP geolocation of active user/presence events as travel indication.

(TS//SI/REL TO USA, FVEY) The DNI analytic operates in one of two modes. The first mode accepts a list of tasked targets via UTT, and attempts to identify co-travelers for those targets that have been deemed to have travelled during a specified time window (typically 30 days). The analytic only considers targets that traveled between at least 2 countries in a given month. For these traveling targets, candidate co-travelers are scored based on how many times they were seen in the same locations during the same times as the target. Target locations are given by DNI selector IP geolocation, provided by ASDF enriched

with GEO reference data (or geo-tagging where available). Because this data provides city-level location resolution, co-traveler candidates are assigned scores based on the extent to which they were seen in the same cities and on the same days as targets.

(TS//SI/REL TO USA, FVEY) The second mode accepts a pattern representing target travel across spanning countries of interest (e.g., ███████████), and optionally, the days on which the countries were visited. In this mode, the TAC/Café/TMAC DNI Co-travel analytic in this mode identifies travelers that (at minimum) match the pattern. All candidates that match the pattern are regarded as possible co-travelers.

 (S//REL TO USA, FVEY) The result of these analytics is a QFD monthly roll-up that can be queried.

Status and Summary

Status	Source Data	Caveats
Available to developers with access to Ghostmachine (GM-PLACE)	- Tasked DNI selectors (UTT) - Geotagged ASDF data - User-provided travel patterns	- Tasked targets or travel patterns provided as input; results include tasked and untasked targets - Analytic operates at the country level to determine travel/city level for co-traveler determination, and designed to provide monthly QFD roll-up - Proxies and other shared IP settings can render IP geolocation susceptible

Future Work
(S//SI/REL TO USA, FVEY) The TAC/Café/TMAC DNI Co-traveler team also considered capabilities to enable follow-on queries utilizing CHALKFUN for convergence efforts to identify roaming handsets as possible DNI target co-travelers.

Other resources
https://ncmd-satc01.ncmd.nsa.ic.gov/gambit/public/q/dni_travel_analytic_cloud_version

https://wiki.nsa.ic.gov/wiki/Cafetravel_dni_co-travelers

TAC/Café/TMAC DNR Co-Traveler Analytic
Café Spin 2 (January – July 2012)

Background

(TS//SI//REL TO USA, FVEY) The Café project involved TMAC, SSG, T1212, and S2I5 working in concert to develop both DNI and DNR cloud-based travel analytics. The absence of a cloud-based solution that could run over bulk data motivated this initiative. The Café objective was to merge the DNI and DNR analytics to create one complete co-travel analytic; however the DNR co-traveler analytic, described below, is currently still under development.

(TS//SI//REL TO USA, FVEY) The DNR cloud-based analytic considers all known targets (tasked in OCTAVE) that have traveled within a given date range (e.g., monthly roll-up to five month range), and attempts to find their co-travelers. Co-travelers are defined as individuals that were seen in the same area (currently defined by VLRs) around the same time as the targets. The output includes both tasked and untasked selectors as possible co-travelers with the tasked seeds. Each possible co-traveler is assigned a score that indicates the probability of co-travel with the seed. Higher scores are assigned to co-travelers that are seen at more of the same locations and closer in time (pairs are given one point if seen within one hour, and a half point if seen within two hours of each other).

Status and Summary

Status	Source Data	Caveats
Analytic has been tested on FASCIA data on GM-PLACE **Command line interface available to developers**	- FASCIA data on GM-PLACE - ~40B rows in the GM PLACE CLOUDBASE table - CHALKFUN Enrichment (VLR Country mapping) - CLOUDBASE Events (IMSI,IMEI) rounded to nearest hour	- Analytic only considers tasked selectors as seeds - Source data provided by VLRs - Co-travel events are rolled-up by the hour

Future Work

(S//SI//REL TO USA, FVEY) Follow-on analysis could take advantage of FASTSCOPE reservation number feature which will return all co-travelers that travel on the same reservation number within a given time period (because reservation numbers are reused, a specific timeframe must be provided).

Other Resources
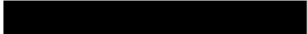
https://wiki.nsa.ic.gov/wiki/DNR_Traveler

https://wiki.nsa.ic.gov/wiki/DNR_Co-Traveler

https://wiki.nsa.ic.gov/wiki/DNR_Travel_Pattern

DNR Co-Traveler Manual Analysis

Taken from: https://ncmd-satc01.ncmd.nsa.ic.gov/gambit/public/q/dnr_co_travel_based_on_similiar_cell_ids_over_a_time_frame

████████████████

1. Start with a target selector (e.g. IMSI)
   2. Query the target selector for PCS events to identify cell towers this target his hitting off of and at what date/time.
   3. Note the cell towers, location of the cell towers, and the date/times
   4. Query those cell towers (and other cell towers in the area) for those dates and times to identify other users who are hitting off of those towers
   5. Compare the results of the users hitting off of the cell towers.
   6. Rank the selectors as being possible candidates for co-travelers based on what cell towers they hit on at the right times.
   7. Selectors that are reliably seen to be hitting off of the same towers at the same times more than others should get a higher rank.

Summary

(S//SI/REL TO USA, FVEY) At the beginning of this paper, we presented a number of key issues and questions. Many of the analytics define themselves by (1) the key issues they address in novel ways and (2) the types of source data on which they operate.

(S//SI/REL TO USA, FVEY) The key issues section highlights capabilities that might improve the accuracy of the analytic results. For example, analytics that have knowledge about the locations of GCIDs and VLRs and can augment their procedures with non-SIGINT data such as geographic and terrestrial data. This information contains knowledge about the locations of highways and roads. Analytics that can geographically validate routes between meeting points can then use this information to constrain the possible co-travel routes and candidate co-travel selectors along those routes.

(S//SI/REL TO USA, FVEY) Analytics that can operate on a variety of different source data formats, including both DNI and DNR, benefit from the ability to exploit divergent data sources to develop more complete pictures of target travel behavior.

(S//SI/REL TO USA, FVEY) The co-travel analytics in this study are at various stages of development, testing, and deployment. One possible way forward could be to have an independent organization[2] perform a formal evaluation of these analytics using a common test dataset. This would enable a fair comparison and assessment of the analytics' processing time, efficiency, and accuracy. Understanding the advantages and challenges of each analytic against a common test dataset with ground truth may facilitate planning for future work.

---

[2] An independent organization is one that is not involved in the development of any of these analytics and that does not have a stake in the outcome.

Acknowledgments

(U//FOUO) Thanks to all of the participants in this study, listed as POCs under the individual analytics. These individuals participated in face-to-face meetings and phone interviews, and provided the details of their analytics to this study through briefings and write-ups. This compilation would not have been possible without the cooperation of these contributors. Special thanks also to the Geospatial Analysis Support Center for their contributions to the section on Issues and Questions.

Summary Table of Co-Travel Analytics

Name of Analytic	Summary	Source Data	Architecture	Status	Caveats
CHALKFUN	Analytic computes the date, time, and network location of any (**tasked or untasked**) mobile phone over some time period, and then looks for other mobile phones that were seen in the same network locations around a one hour time window. When a selector was seen at the same location (e.g., VLR) during the time window, the algorithm will reduce processing time by choosing a few events to match over the time period. Chalkfun is SPCMA enabled.	- All FASCIA data containing VLR and GCID information	- Cloud-based version could be available as early as September 2012.	- Operational; Available at analysts desktops	- Current version is not cloud-based and can have long processing times, however cloud-based solution is imminent. - Analytic will only return co-travelers on the same provider network
DSD Co-Travel Analytic	Predicts target locations and co-travelers by calculating time-based travel trajectories and identifying likely path intersections between observed locations. The analytic calculates travel times at waypoints similar to that used in turn-by-turn navigation systems.	-Mobile CDRs	- Netezza - Could be implemented in Cloud-based architecture (Hadoop/ MapReduce or Accumulo)	- Implemented and tested at DSD	- Requires Netezza (current implementation) - Requires Renoir
Geospatial Analysis Tradecraft Center	Determines whether two entities (e.g. devices) could have been **co-located** by	- Geohashes of GPS point event data.	Cloud-based	Prototype service implemented on NGANet. Not yet	- Requires event locations and times for every selector. - Designed for 1 km grid-based

Name of Analytic	Summary	Source Data	Architecture	Status	Caveats
(GATC) Opportunity Volume Analytic	considering the possibility of their travel paths intersecting. Computes possible travel routes for each entity between specified events, considering terrain, land cover, and road network data.			ported to NSANet.	locations and 15 minute time intervals. - Co-travel capability would require analyst to define a series of meetings at specified locations.
▮▮▮▮ Co-Traveler Analytic	The analytic computes event sequences of LAT, LONG, and time for each tasked selector. These are called "tracks". Each selector's tracks are pairwise-compared to the others and a measure of similarity in time and space is computed. The analytic works cross-network by computing target "closeness" based on the GCID Lat/Long GEO information and time.	- Sortinglead summaries of FASCIA data on GM-PLACE and GM-▮▮▮▮ - RT-RG regional GSM collection ▮▮▮	Cloud-based	Initial development completed. In testing phase, not yet operational	- This cloud analytic is oriented to work on 7 to 30 days worth of regional collection. - Analytic only considers tasked selectors as seeds. - Analytic does not consider targets that do not travel outside a 20 to 50 km radius. - Track dataset must be repopulated for each data update
▮▮▮▮ Co-Traveler Analytics	- The Fast Follower (FF) analytic considers non-SIGINT reference paths for station personnel based on detailed knowledge of the entity's location. Candidate followers are determined by identifying other individuals whose path matches the reference path in space and time. - The Meet&Greet Spatial Chaining (MGSC) analytic	- Smartphone data from ELKPRINTS - Reference-path data (FF) - List of selectors (MGSC)	Cloud-based Implemented in Java and ported to MapReduce	The MGSC analytics has been tested on real ELKPRINTS data, but results have not been validated by operational analysts. The FF analytic has been tested on made-up data.	- Analytic designed for precise geolocation data (e.g., from smartphones) - MGSC analytic would require the analyst to define a series of meetings

Name of Analytic	Summary	Source Data	Architecture	Status	Caveats
	applies a maximum velocity movement parameter to approximate the time that an individual was at each location. Co-travelers are identified by discovering pairs of selectors that meet duration and distance thresholds set by the analyst.				
PACT NGA-NSA GATC Analytic	Identifies clusters of waypoints for each target handset. Identifies which pairs of handsets contain similar waypoint clusters. Pairs are scored based on the number of waypoint clusters that match.	- ▓▓▓▓ data from CULTWEAVE via ICReach (e.g. ~5M locations over 6 years for ▓▓▓ 200K locations per day)	Cloud-based Hadoop MapReduce framework	Tested on▓▓▓▓ data from CULTWEAVE. Patterns stored in QFD. In process of transitioning PACT to NSA/S2.	- Analytic designed for ▓▓▓▓▓ point data
R6 SORTINGLEAD Co-Traveler Analytic	Analytic accepts a **tasked or untasked** selector and timeframe as input, and then derives an itinerary for the selector that includes the CELL IDs and/or VLRs. The itinerary is based on a series of waypoints. The analytic searches for other selectors that were "near" these waypoints in space and time. Candidates are scored and prioritized.	- In testing phase to be replacement back-end for the current production CHALKFUN co-traveler tool	Cloud-based MapReduce	- FASCIA PCS Sortinglead Summaries	- Analytic cannot recover cross-network co-travelers - Analytic will not be effective against stationary (non-traveling) targets - Processing is memory intensive - Analytic is sensitive to large cells, VLRs, and dense areas - Not directly applicable to sat phones with LAT/LONG information - Results can be sensitive to timeframe chosen as input

Name of Analytic	Summary	Source Data	Architecture	Status	Caveats
					(not effective for large queries across multiple countries and large time frames)
RT-RG Sidekicks	(TS//SI/REL TO USA, FVEY). This analytic computes "movement summaries" of **tasked** selectors. These are lists of locations that a target visited during the timeframe of interest. Then, the system discovers pairs of targets that could be traveling together by comparing their movement summaries, factoring out pairs that could not have reasonably arrived at the meeting waypoints within 10 minutes of each other. Because this analytic considers physical (LAT/LONG) locations and travel velocities, it can provide co-traveler results that include selectors on different provider networks.	- Currently running on RT-RG ▇▇▇▇▇▇ - Could possibly scale to FASCIA event summaries	Cloud-based	- QFD available at RT-RG analyst desktop. - RT-RG Tools: Goldminer, CHET, GEOT	- Requires pre-computing all selectors against all selectors, which can be expensive - Current output includes only tasked selectors - Analytic is not designed for stationary targets
Scalable Analytics Tradecraft Center (SATC) Geospatial Lifelines Co-Travel QFD	This QFD first generates geohashes using GSM event data, and then calculates transition lines indicating that a device traveled from one geohash to another. The likeliness of co-travel is based on dwell times at travel	- Geohashes of GSM event data retrieved from FASCIA.		Analytic tested on 90 days of GSM event data from ▇▇▇▇▇▇ ▇▇ Code is available through SATC, but analytic is no	- Analytic designed for GSM data, but could be applied to other types of data - Oriented to targets that remain in one location for at least 2 hours - Requires Geocoded source data for generating

Name of Analytic	Summary	Source Data	Architecture	Status	Caveats
	endpoints, and the following measurements: net distance, time of transition (mins), speed (kph), Azimuth, and number of travel segments.			longer under development.	Geohashes
SSG Common IMSIs Analytic	This SEDB JEMA model finds SIM card activity seen on cell tower panels in multiple areas. The analyst inputs areas of interest and time range. The analytic returns an excel file with a list of IMSIs seen in those areas at that time, enriched with OCTAVE tasking information.	OCTAVE and FASCIA data	Tower QFD	Operational, available in JEMA.	- Cell tower locations in OCTSKYWARD can be imprecise. - The SEDB Tower QFD summarizes IMSIs by LAIC by day. - Summaries by MSISDN or IMEI are not available.
Target Analysis Center (TAC)/Café/ Travel and Mobility Analysis Center (TMAC) DNI Co-Travel Analytic	Discovers candidate co-travelers based on how many times selectors were seen in the same countries and cities during the same months as tasked targets. Locations are given by DNI selector IP geolocation, provided by ASDF enriched with GEO reference data.	- Tasked DNI selectors (UTT) - Geotagged ASDF data - User-provided travel patterns	Cloud-based GM-PLACE	Available to developers with access to Ghostmachine (GM-PLACE)	- Tasked targets provided as input; results include tasked and untasked targets - Analytic operates at the country level, and designed to provide monthly QFD roll-up - Proxies can make IP resolution challenging
TAC/Café/ TMAC DNR Co-Traveler Analytic	(TS//SI/REL TO USA, FVEY) The DNR cloud-based analytic considers all known targets (**tasked** in OCTAVE) that have traveled within a given month, and attempts to find their co-travelers. Co-travelers are	- FASCIA data on Ghostmachine - 40.7B rows in the CLOUDBASE table - CHALKFUN Enrichment (VLR	Cloud-based GM-PLACE	Under development	- Analytic only considers tasked selectors as seeds.

Name of Analytic	Summary	Source Data	Architecture	Status	Caveats
	defined as individuals that were seen in the same area (defined by Country, VLR, or Cell ID) around the same time as the targets. The output includes both **tasked and untasked** selectors as possible co-travelers with the tasked seeds.	Country mapping) - CLOUDBASE Events (IMSI,IMEI) rounded to nearest hour			