Relazione visual analytics

1.0 - Introduzione

VAST Challenge 2018 propone uno scenario fittizio in cui operare tramite tecniche di visual analytics. Questo il contesto: Mistford è una città di medie dimensioni situata a sud-ovest del Boonsong Lekagul Wildlife Preserve. La città ha una piccola area industriale con quattro attività di produzione leggera. Mistford e la riserva faunistica sono alle prese con un fenomeno di decimazione della popolazione del Pipit azzurro, un uccello localmente molto amato che nidifica nella zone boschive limitrofe alla città. Le coppie nidificanti dell'uccello sembrano essere diminuite in modo allarmante, provocando l'avvio di un'indagine che ha coinvolto Kasios Office Furniture, un'azienda manifatturiera di Mistford che sembra essere implicata nella scomparsa del Pipit azzurro.

La challenge si divide in due fasi:

1. Vengono forniti dati relativi alla concentrazione di sostanze tossiche in ciascuna area di interesse. L'obiettivo è quello di valutare possibili impatti negativi sull'habitat del Pipit, specie estremamente suscettibile a certi tipi di molecole.

2. Vengono forniti dati relativi alle attività dell'azienda Kasios Office Furniture al fine di identificare singoli o gruppi di dipendenti che potrebbero aver causato danni ambientali con condotte fraudolente e scorrette.

2.1 - Descrizione dati challenge 1

Come accennato in precedenza, per affrontare questa prima challenge vengono forniti valori continui in relazione alla quantità di 106 sostanze (non obbligatoriamente tossiche) disciolte nelle falde acquifere in 10 località per un range temporale di 20 anni. Un'istanza ha questo formato:

ID: 2221
Value: 9.1
Location: Boonsri
Sample date: 11-Jan-98
Measure: Water temperature

Il seguente schema riassume i valori che le variabili categoriche (Location e Measure) o discrete (Sample date) possono assumere:

Misura	Valori possibili
Location	Chai, Kannika, Chai, Kohsoom, Somchair, Boonsri, Sakda, Busarakhan, Tansanee, Achara, Decha
Measure	1,2,3-Trichlorobenzene, 1,2,4-Trichlorobenzene, Acenaphthene, Acenaphthylene, AGOC-3A, Alachlor, Aldrin, alpha-Hexachlorocyclohexane, Aluminium, Ammonium, Anionic active surfactants, Anthracene, AOX, Arsenic, Atrazine, Barium, Benzo(a)anthracene, Benzo(a)pyrene, Benzo(b)fluoranthene, Benzo(g,h,i)perylene, Benzo(k)fluoranthene, Berilium, beta-Hexaxchlorocyclohexane, Bicarbonates, Biochemical, Oxygen, Boron, Cadmium, Calcium, Carbonates, Cesium, Chemical Oxygen Demand (Cr), Chemical Oxygen Demand (Mn), Chlorides, Chlorodinine, Chromium, Chrysene, Copper, Cyanides, Dieldrin, Dissolved organic carbon, Dissolved oxygen, Dissolved silicates, Endosulfan (alpha), Endosulfan (beta), Endrin, Fecal coliforms, Fecal streptococci, Fluoranthene, Fluorene, gamma-Hexachlorocyclohexane, Heptachlor, Heptachloroepoxide, Hexachlorobenzene, Indeno(1,2,3-c,d)pyrene, Inorganic nitrogen, Iron, Isodrin, Lead, Macrozoobenthos, Magnesium, Manganese, Mercury, Methoxychlor, Methylosmoline, Metolachlor, Naphthalene, Nickel, Nitrates, Nitrites, Organic nitrogen, Orthophosphate-phosphorus, Oxygen saturation, p,p-DDD, p,p-DDE, p,p-DDT, PAHs, PCB101, PCB 118, PCB 138, PCB 153, PCB 180, PCB 28, PCB 52, Pentachlorobenzene, Petroleum hydrocarbons, Phenanthrene, Potassium, Pyrene, Selenium, Silica (SiO2), Simazine, Sodium, Sulfides, Sulphates, Tetrachloromethane, Total coliforms, Total dissolved phosphorus, Total dissolved salts, Total extractable matter, Total hardness, Total nitrogen, Total organic carbon, Total phosphorus, Trifluralin, Water temperature, Zinc.
Sample date	11/01/1998 - 30/12/2016

Per completare questa prima parte della challenge è stato necessario rispondere, attraverso le analisi svolte sui grafici, ad alcune domande specifiche:

1) Caratterizza la situazione passata e recente per quanto riguarda la contaminazione chimica nelle vie navigabili Boonsong Lekagul. Vedi qualche tendenza di possibile interesse in questa indagine?.
2) Quali anomalie trovi nel set di dati dei campioni della via navigabile? In che modo influiscono sull'analisi dei potenziali problemi per l'ambiente? Il Dipartimento di idrologia sta raccogliendo dati sufficienti per comprendere la situazione globale in tutta la Riserva? Quali modifiche proporresti di apportare all'approccio di campionamento per comprendere meglio la situazione?

3) Dopo aver esaminato i dati, alcune delle tue scoperte causano particolare preoccupazione per il Pipit o altri animali selvatici? Suggeriresti qualche cambiamento nella strategia di campionamento per comprendere meglio la situazione dei corsi d'acqua nella Riserva?

2.3 - Stato dell’arte

Altri partecipanti alla challenge hanno dato il loro contributo nel tentativo di rispondere alle domande riportate nel capitolo 2.1.
Per rispondere alla prima domanda è ragionevole caratterizzare la situazione passata e recente con un sottoinsieme rappresentativo dei dati. L'analisi visiva può essere utilizzata per scegliere quel sottoinsieme. Alcune i valori di alcune sostanze vengono misurati frequentemente e in molte posizioni, mentre altri vengono misurati in poche posizioni in un breve periodo di tempo. Molte sostanze chimiche hanno troppe poche letture per comprovare eventuali cambiamenti nel tempo.
Sono stati impiegati istogrammi di complessità variabile che mostrano il numero di misurazioni per ciascuna sostanza nel corso del tempo. Gli istogrammi più complessi sfruttano scale cromatiche categoriche per identificare la località in cui sono stati registrati i dati. Vengono impiegati anche boxplot (pessimi per una visualizzazione d’insieme perché troppo ingombranti) e bubble chart, scelta opinabile in quanto si associa al count dei valori continui una visualizzazione basata su aree. Fanno eccezione i casi estremi, che ottengono maggiore risalto perchè descritti da circonferenze molto ampie o molto ridotte.
I trends delle unità di misura rilevanti sono stati visualizzati tramite line chart, permettendo il confronto fra componenti differenti nello stesso range temporale e per le stesse località.
Al fine di valutare perturbazioni rispetto alla media storica di ciascuna sostanza chimica per ciascun luogo, sono state sviluppate in rari casi delle heatmap a partire dalla cartina fornita insieme ai dati. Più spesso sono stati impiegati semplici line chart.
Infine, per identificare le anomalie presenti nei dati, sono stati impiegati nuovamente istogrammi che mostravano il count delle misurazioni nel tempo.

2.4 - Dashboard e interactive line chart

La prima challenge è stata risolta sviluppando 2 componenti: Dashboard e interactive line chart. La dashboard si compone di 4 widgets che condividono interattivamente i dati e permettono di fissare il panorama delle misurazioni distribuite per località, misura e anno.
Il primo widget è una heat map che descrive il count delle misurazioni totali diviso per anno e location. Il secondo è un bar chart che mostra il numero totale di misurazioni divise per location. Il terzo è anch’esso un bar chart, ma mostra il numero totale di misurazioni effettuate divise per tipologia di misurazione. L’ultimo mostra il numero di misurazioni effettuate divise per anno.

Il componente è costruito per spingere l'utente ad esplorare i dati. I colori, tranne che per la heat map, vengono usati per aiutare nella selezione dei dati e non sono correlati con le misure.
L' interactive line chart permette all'utente di visualizzare i dati in maniera dettagliata, muovendosi fra le dimensioni tempo (slider), spazio e tipologia di misurazione(selettori).
Il design minimale e pulito non affolla lo schermo e risulta molto intuitivo nell'utilizzo.

2.5 - Risultati dell'analisi

Vengono fornite le risposte ai quesiti della challenge facendo riferimento ai grafici:

1) Caratterizza la situazione passata e recente per quanto riguarda la contaminazione chimica nelle vie navigabili Boonsong Lekagul. Vedi qualche tendenza di possibile interesse in questa indagine?

Tramite l'interactive line chart sono state identificate due pericolose tendenze, forse collegate, nelle regioni di Chai (Immagine 1-2), Kohsoom (Immagine 3-4), Somchair(Immagine 5-6). Ad un incremento notevole della Methylosmolina (sostanza inquinante pericolosa per il Pipit) tra la fine del 2015 e l'inizio del 2016, corrisponde un brusco decremento della Chlorodinina. Questo trend potrebbe identificare un'attività sistematica di sversamento della sostanza chimica nelle falde acquifere delle due località.

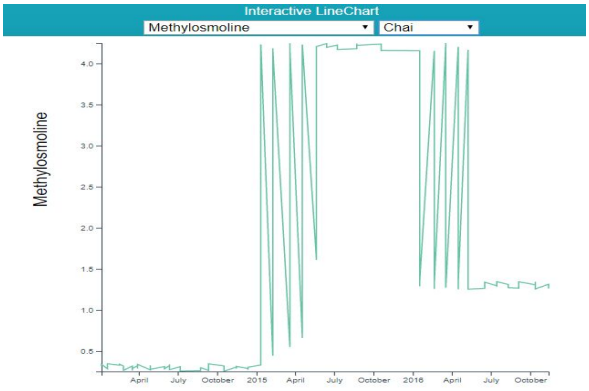

Immagine 1 - Methylosmolina a Chai

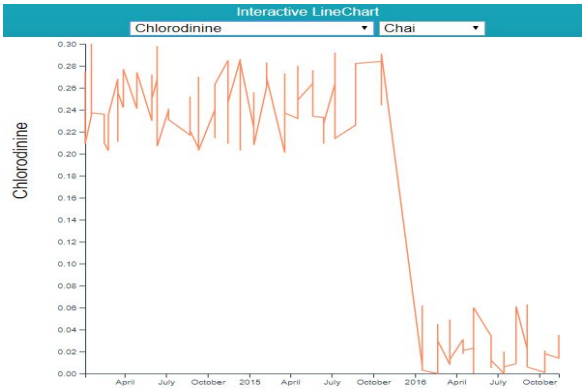

Immagine 2 - Chlorodinina a Chai

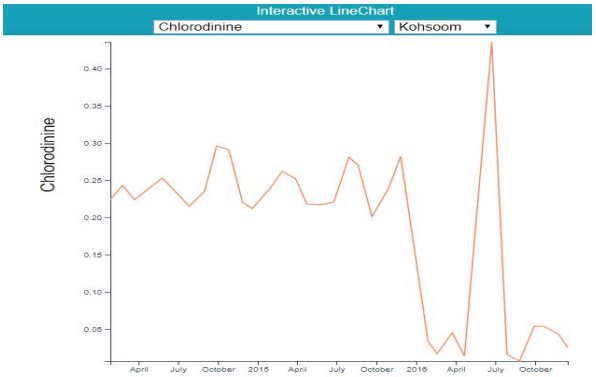

Immagine 3 - Methylosmolina a Kohsoom

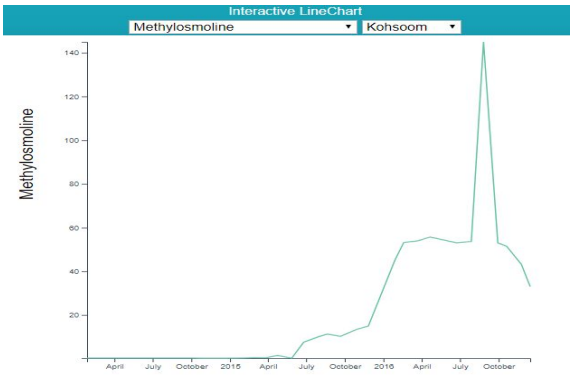

Immagine 4 - Chlorodinina a Kohsoom

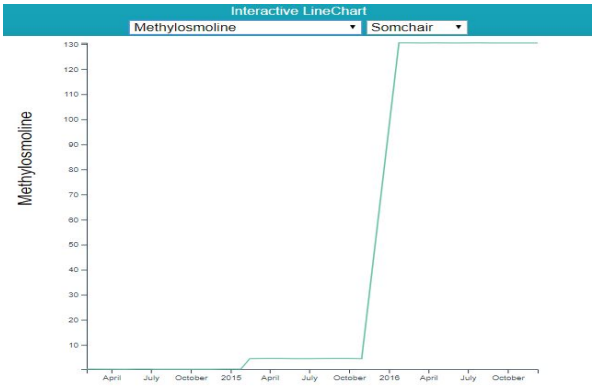

Immagine 5 - Methylosmolina a Somchair

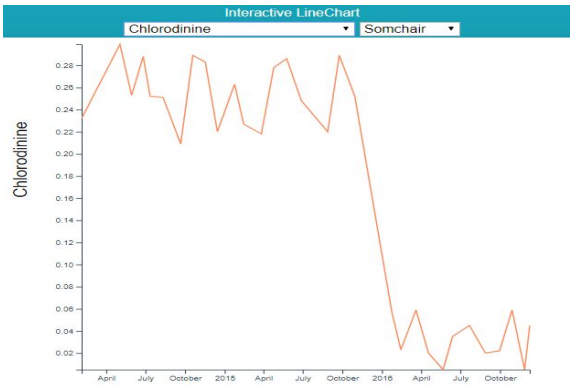

Immagine 6 - Chlorodinina a Somchair

2) Quali anomalie trovi nel set di dati dei campioni della via navigabile? In che modo influiscono sull'analisi dei potenziali problemi per l'ambiente? Il Dipartimento di idrologia sta raccogliendo dati sufficienti per comprendere la situazione globale in tutta la Riserva? Quali modifiche proporresti di apportare all'approccio di campionamento per comprendere meglio la situazione?

Dalla dashboard appare chiaro che soltanto le località di Sakda (20k misurazioni), Kannika(più di 20k misurazioni), Chai (più di 28k misurazioni) e Boonsri (circa 28k misurazioni) hanno set di misurazioni sufficientemente cospicue da permettere un' analisi.

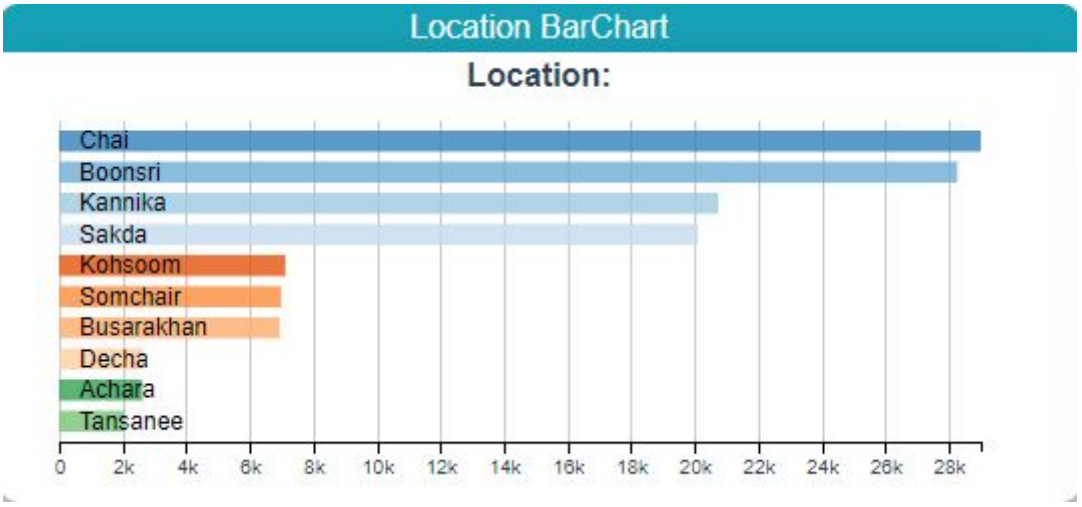

Immagine 7 - Location bar chart - Count delle misurazioni totali

Inoltre, misurazioni di una certa rilevanza come la Methylosmolina, la AGOC-3A e la Chloridina sono estremamente rare.

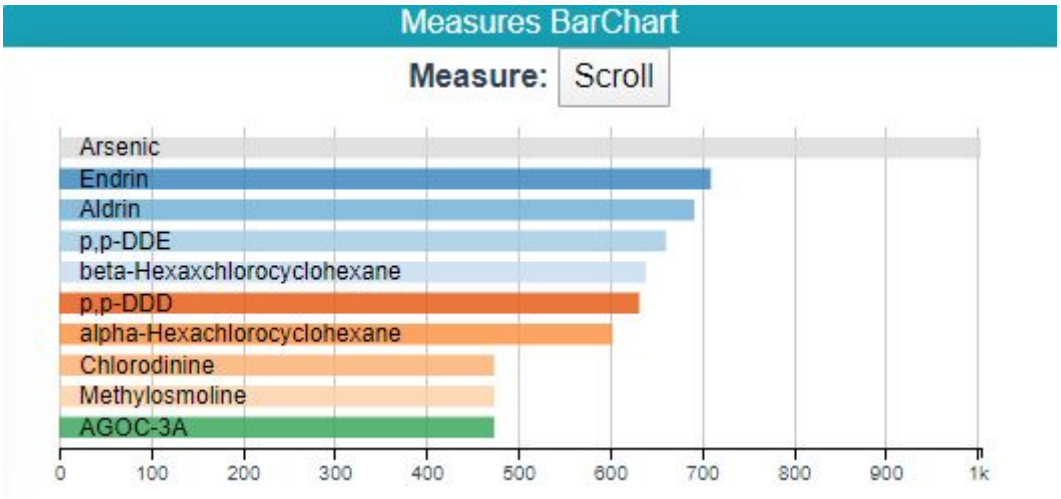

Immagine 8 - Measures bar chart - Count delle misurazioni di interesse

Attraverso l'interactive line chart è possibile identificare un'anomalia nella registrazione della misura water temperature nella località Chai. Ciò potrebbe indicare un errore nell'acquisizione dei dati o un malfunzionamento della strumentazione atta ad acquisirli.

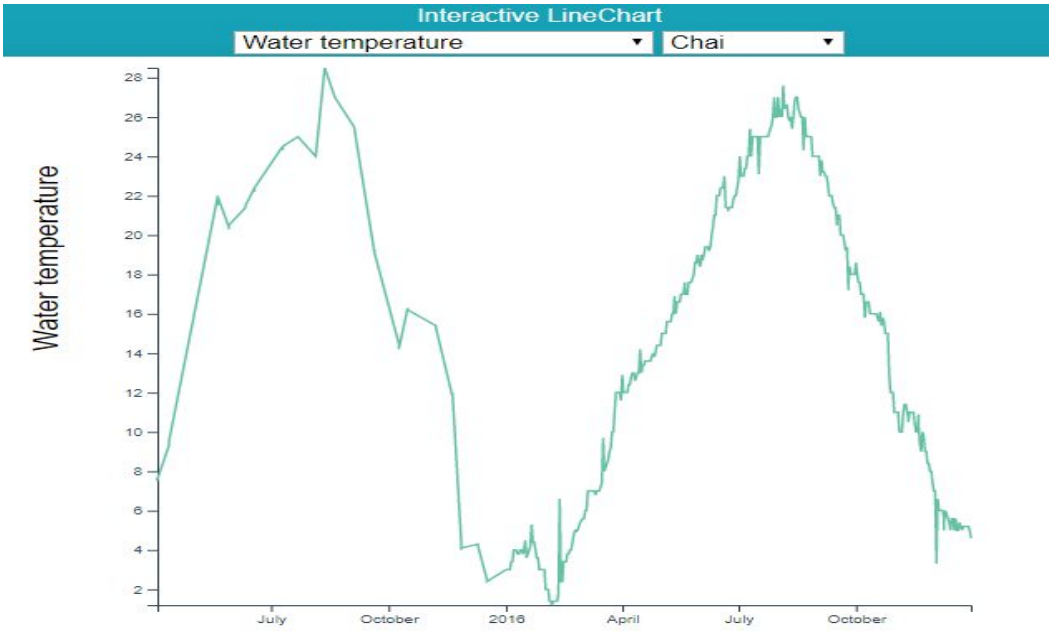

Immagine 9 - Interactive Line Chart - Anomalia water temperature Chai

3) Dopo aver esaminato i dati, alcune delle tue scoperte causano particolare preoccupazione per il Pipit o altri animali selvatici? Suggeriresti qualche cambiamento nella strategia di campionamento per comprendere meglio la situazione dei corsi d'acqua nella Riserva?

La bash board rivela che alcune posizioni hanno pochissime misurazioni chimiche (Immagine 10). Sulla base di questo, si raccomanda di aumentare i test a Tansanee, Decha

e Achara. Fortunatamente, questi tre siti non sono vicini al sito di dumping sospetto. Tuttavia, anche Somchair e Kohsoom hanno livelli di misurazione relativamente bassi. Entrambi questi siti mostrano prove di contaminazione con Methylosmolina, quindi i test dovrebbero essere aumentati anche in questi luoghi.

Immagine 10 - Misurazioni Methylosmolina, AGOC-3A e Chloridina in Chai e Boonsri (sinistra) e in Somchair e Kohsoom (destra)

3.1 - Descrizione dati challenge 2

Come accennato nel capitolo 1.0, per affrontare la seconda parte della challenge vengono messi a disposizione dati relativi all'attività dell'azienda Kasios Office Furniture, al fine di identificare singoli dipendenti che potrebbero aver causato danni ambientali con comportamenti fraudolenti e scorretti.
Vengono forniti 8 csv contenenti record di attività: chiamate, e-mail, acquisti e riunioni. I dati includono solo l'origine di ciascuna transazione, il destinatario e l'ora della transazione. I contenuti delle e-mail o delle telefonate non sono disponibili. Esiste anche un indice aziendale che mostra il nome di tutti i membri dell'azienda e il numero ID associato. Ci sono 642.631 individui nell'indice. Quattro file di dati (uno per ciascun tipo di transazione) coprono le attività dell'intera azienda. Altri quattro file, anch'essi divisi come i primi, contengono solo le attività sospette. Le istanze si presentano con questo formato:

Source: 3029
Etype: 1
Destination: 2383
Time stamp: 3452352345

Il seguente schema riassume i valori che le variabili categoriche (Etype) o discrete (Source, Destination e Time stamp) possono assumere:

Misure	Valori possibili
Source	Id del dipendente nella compagnia da cui parte l'email o la chiamata, che organizza il meeting o esegue l'acquisto
Etype	0 per le chiamate 1 per le email 2 per gli acquisti 3 per i meeting
Destination	Id del dipendente nella compagnia a cui è destinata l'email o la chiamata, che è invitato al meeting o che riceve l'oggetto acquistato.
Time stamp	Data dell'evento in secondi a partire dall'11 maggio 2015 alle 14:00

Per completare questa seconda parte della challenge è stato necessario rispondere, attraverso le analisi svolte sui grafici, ad alcune domande specifiche:

1) Utilizzando i quattro grandi set di dati di Kasios International, combinare le diverse fonti per creare una singola immagine dell'azienda. Caratterizza i cambiamenti nell'azienda nel tempo. Secondo le comunicazioni e le abitudini di acquisto, l'azienda sta crescendo?
2) Combina le quattro origini dati per gruppo che l'insider ha identificato come sospetto e individua il gruppo nel set di dati più grande. Determina se qualcun altro sembra essere strettamente associato a questo gruppo. Evidenzia quali dipendenti stanno effettuando acquisti sospetti, secondo i dati dell'insider.
3) Utilizzando il gruppo combinato di sospetti creato nella domanda 2, mostra le interazioni all'interno del gruppo nel tempo.
    a) Caratterizza la struttura organizzativa del gruppo e mostra un quadro completo delle comunicazioni all'interno del gruppo.
    b) La composizione del gruppo cambia nel corso delle loro attività?
    c) Come cambiano le interazioni del gruppo nel tempo?
4) L'insider ha fornito un elenco di acquisti che potrebbero indicare attività illecite in altre parti dell'azienda. Utilizzando la struttura del primo gruppo annotato dall'insider come modello, è possibile trovare altri casi di attività sospette nell'azienda? Ci sono altri gruppi che hanno struttura e attività simili a questo? Loro chi sono? Ciascuno degli acquisti sospetti potrebbe essere un punto di partenza per la tua ricerca. Fornisci esempi di altri due gruppi che ritieni sospetti e confronta la loro struttura con la struttura del primo gruppo. Le strutture dovrebbero essere presentate come temporali e non solo strutturali (cioè, la sequenza di eventi - A è seguita da B uno o due giorni dopo - sarà importante).

3.3 - Stato dell’arte

Altri partecipanti alla challenge hanno dato il loro contributo nel tentativo di rispondere alle domande riportate nel capitolo 3.1. La soluzione proposta consiste sempre in uno o più grafici della rete sociale sottesa al funzionamento dell’azienda. Questa tipologia di grafico è stata ampiamente testata in molte possibili varianti aggiungendo la selezione dei nodi per isolarne l’ego network, impiegando sliders per le date e per la profondità della rete.

3.4 - Line chart e network graph

La challenge è stata risolta impiegando due componenti: line chart e network graph.
Il primo, simile a quello impiegato nella prima challenge, permette di visualizzare l’andamento dell’azienda nel tempo.
Il network graph permette di visualizzare la struttura criminale all’interno dell’azienda e di tracciarla nel tempo. La visualizzazione si alterna fra sospetti certi e non, permettendo di collocare la struttura criminale in un più ampio quadro aziendale. E’ possibile selezionare ciascun nodo e isolarne l’ego network. I colori usati per i nodi sono blu (non sospetti) e azzurro(sospetti). I link hanno colore in base al tipo di interazione: email arancione, chiamata blu, acquisto rosso e meeting verde.

3.6 - Risultati dell’analisi

1) Utilizzando i quattro grandi set di dati di Kasios International, combinare le diverse fonti per creare una singola immagine dell'azienda. Caratterizza i cambiamenti nell'azienda nel tempo. Secondo le comunicazioni e le abitudini di acquisto, l'azienda sta crescendo?

Il line chart mostra un forte incremento dei meeting, sintomo di grande attività da parte dell’azienda.

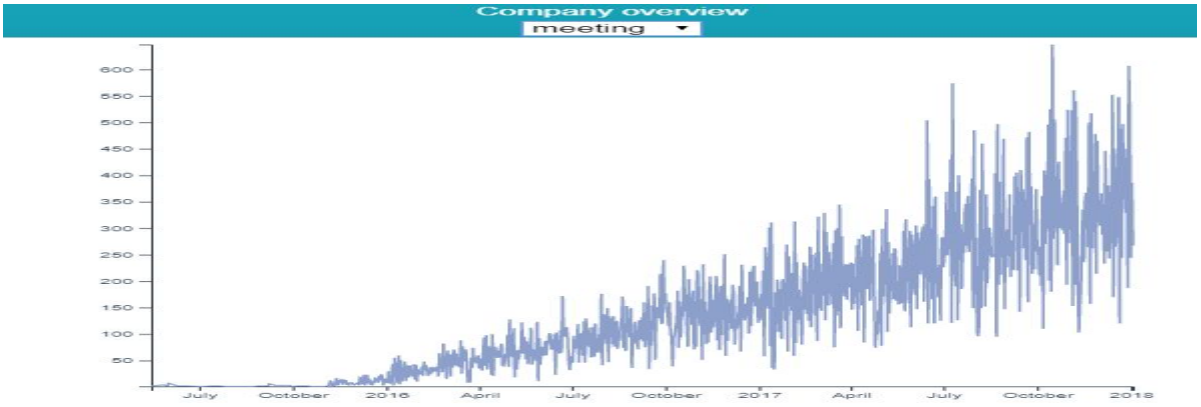

Immagine 11 - Meeting trend

2) Combina le quattro origini dati per gruppo che l'insider ha identificato come sospetto e individua il gruppo nel set di dati più grande. Determina se qualcun altro sembra essere strettamente associato a questo gruppo. Evidenzia quali dipendenti stanno effettuando acquisti sospetti, secondo i dati dell'insider.

I sospetti sono evidenziati nel grafico come nodi azzurri ed è facile intuire la gerarchia dell'organizzazione dal numero di contatti con dipendenti apparentemente non coinvolti (nodi blu). Gli acquisti sono evidenziati in rosso (Immagine 12).

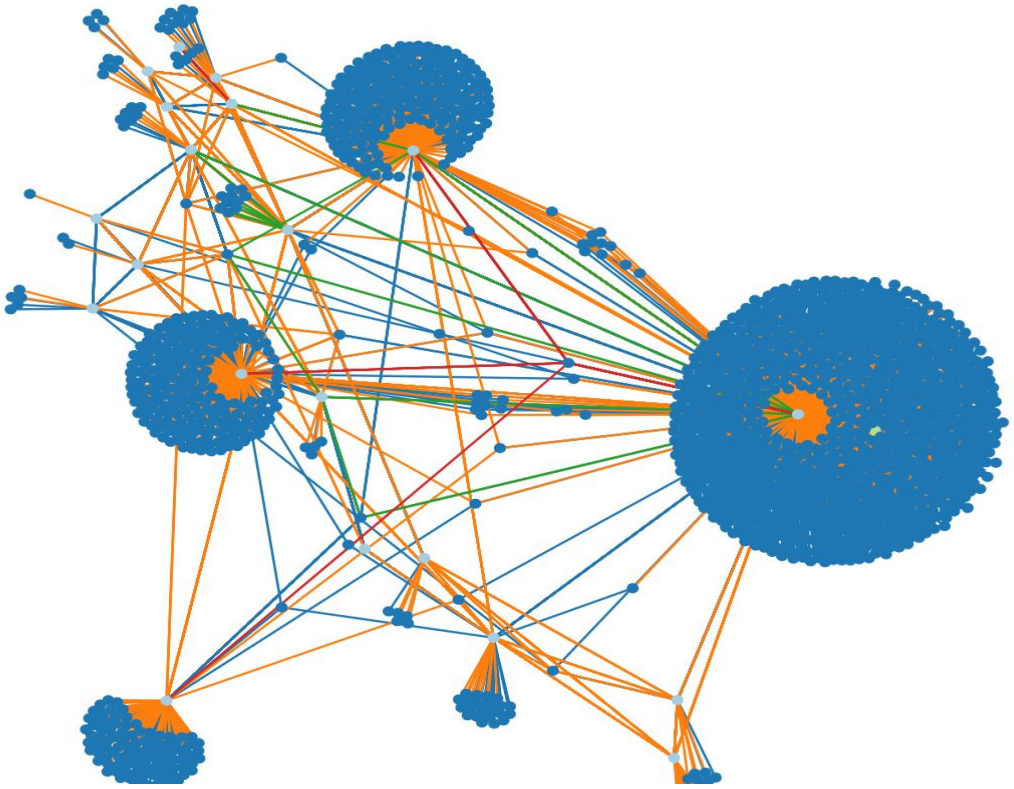

Immagine 12 - Organizzazione e possibili complici

3) Utilizzando il gruppo combinato di sospetti creato nella domanda 2, mostra le interazioni all'interno del gruppo nel tempo.
    a) Caratterizza la struttura organizzativa del gruppo e mostra un quadro completo delle comunicazioni all'interno del gruppo.
    b) La composizione del gruppo cambia nel corso delle loro attività?
    c) Come cambiano le interazioni del gruppo nel tempo?

Nell'immagine 12 possiamo notare come la struttura criminale sia organizzata e gerarchicamente ben definita. Un'ipotesi potrebbe essere che ciascuno dei componenti di rilievo mantiene contatti con una cerchia di possibili sospetti in base all'attività che svolge all'interno dell'organizzazione.
Un altro elemento che caratterizza l'organizzazione è l'impiego di intermediari nelle comunicazioni fra i leader dell'organizzazione. L'immagine 13 lo mette in evidenza.

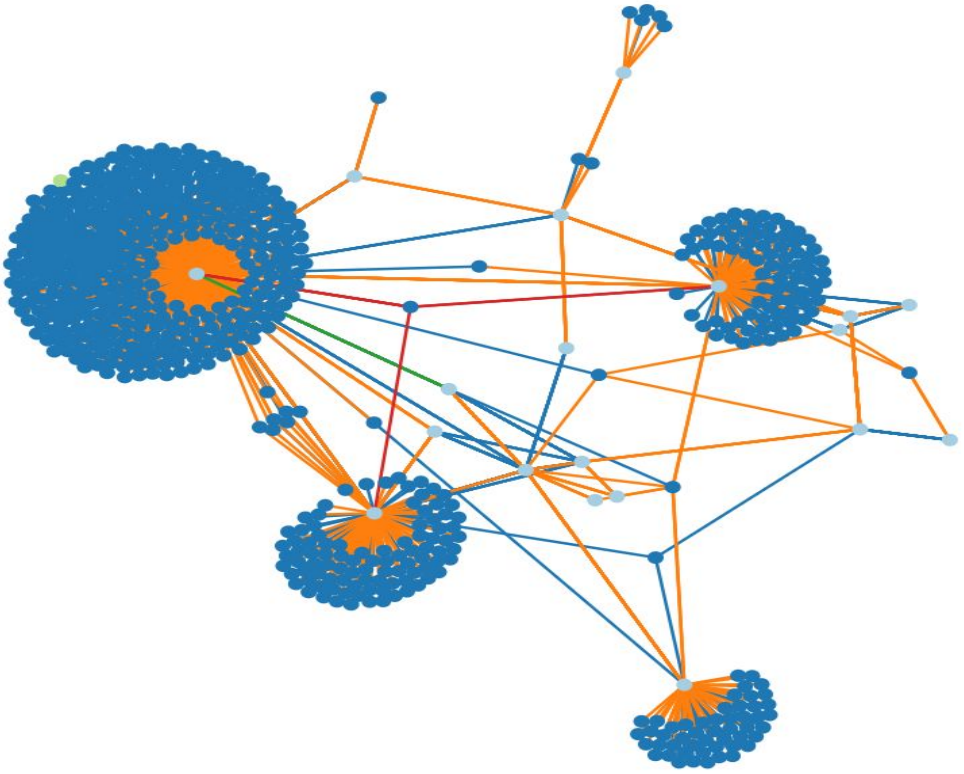

Immagine 13 - Intermediari

E’ inoltre possibile tracciare i mutamenti nell’organizzazione usando lo slider per scoprire che elementi come Ramiro Gualt, Refugio Orrantia, Glen Grant e Dylan Ballard hanno abbandonato l’organizzazione.