**Analisi sul dataset sulle relazioni dei Pinguini dell’acquario di Kyoto**

**Studenti:**

*- Ciardi Guido (7090798) (*[*guido.ciardi@stud.unifi.it*](mailto:guido.ciardi@stud.unifi.it)*)*

*- Martinelli Giuseppe (7093926) (giuseppe.martinelli@stud.unifi.it)*

*- Sirgiovanni Massimiliano (7077251) (*[*massimiliano.sirgiovanni@stud.unifi.it*](mailto:massimiliano.sirgiovanni@stud.unifi.it)*)*

*- Tosi Federica (7093967) (*[*federica.tosi@stud.unifi.it*](mailto:federica.tosi@stud.unifi.it)*)*

***Link al dataset****:* [*https://bitbucket.org/mhfeng/penguins\_of\_kyoto/src/master/*](https://bitbucket.org/mhfeng/penguins_of_kyoto/src/master/)

**Descrizione del dataset**

Per questo progetto è stato analizzato un dataset relativo a **59 pinguini** che si trovano nell’**acquario di Kyoto**, distinti tra di loro mediante i loro nomi.

I files presi in considerazione fanno riferimento a **relazioni di diverso tipo** tra i pinguini:

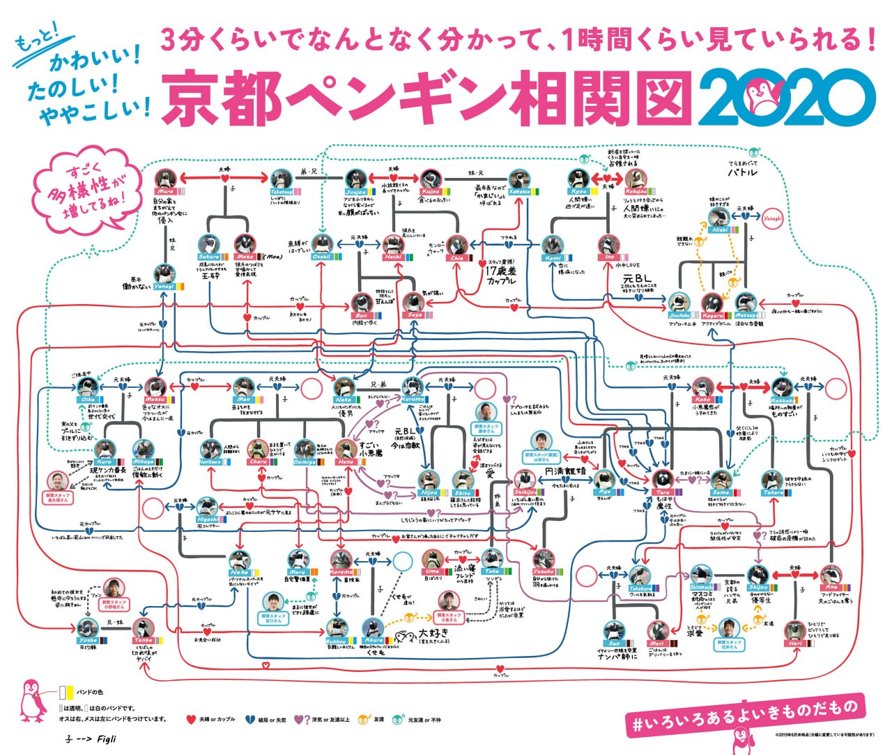
* ***Coppie:*** Relazioni in corso/stabili tra pinguini;
* ***Ex:*** Relazioni passate o rifiutate dai pinguini;
* ***Famiglia:*** Relazioni familiari;
* ***Relazioni complicate:*** Situazioni in cui i pinguini non si trovano in una relazione effettiva *(non sono una coppia)*, ma in certi momenti si relazionano con una logica abbastanza casuale;
* ***Amicizia:*** Relazioni di amicizia;
* ***Nemici:*** Relazioni di inimicizia.

Inizialmente erano presenti più nomi o soprannomi associati ad alcuni pinguini.

Per questo, con un lavoro di **preprocessing,** sono stati rilevati i pinguini che presentavano questa caratteristica e gli sono stati applicati **nomi univoci**.

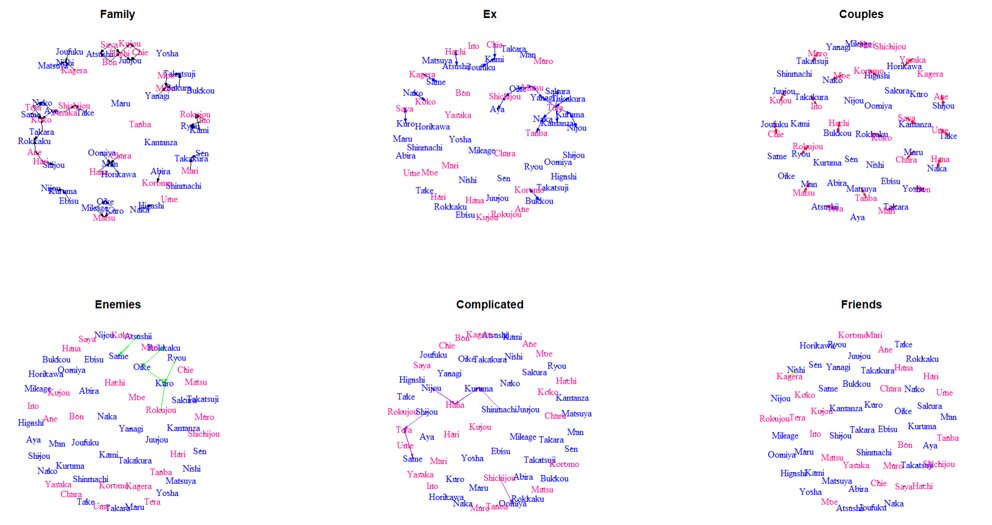
Avendo trovato poi, *sul sito dell'acquario di Kyoto*, una rappresentazione grafica che illustrava il genere e le relazioni tra i vari pinguini, abbiamo pensato di inserire questo attributo manualmente nei dati.

Il grafico su cui ci si è basati è il seguente:



Tutte le relazioni messe a disposizione erano inizialmente rappresentate nei **file csv** trovati sul sito sotto forma di matrici di adiacenza e sono state poi definite come grafi mediante il comando di R *“graph\_from\_adjacency\_matrix”*.

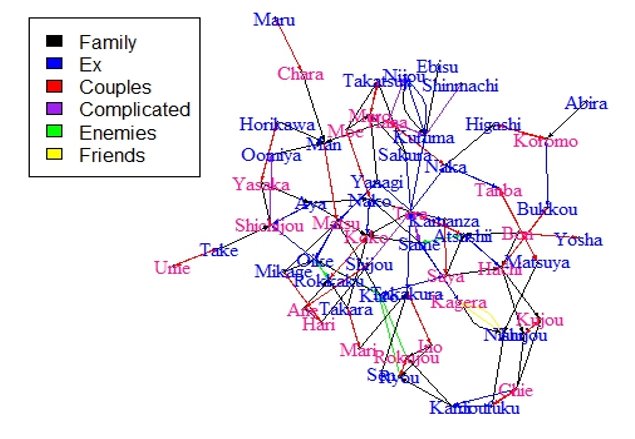
*Sono stati ricavati i seguenti grafi*:



Qui ogni nodo corrisponde a un pinguino, che è rappresentato mediante il suo nome. I **pinguini maschi** sono stati **colorati in blu** e le **femmine in rosa** per evidenziare meglio situazioni più interessanti, come quella del grafo delle coppie, in cui ci sono solo relazioni con coppie di nodi o nodi singoli. Questa considerazione tornerà più avanti in diversi risultati delle analisi che verranno fatte.  
Una caratteristica evidenziata particolarmente da questi grafi è che sono **molto sparsi**, anche se alcuni, come quelli della famiglia, degli ex e delle coppie, presentano più relazioni rispetto ad altri .

Per effettuare le varie analisi sono stati scelti **i primi tre dataset** *(sulla prima riga nell’immagine sopra)*, che sono quelli meno sparsi.

Successivamente, sono stati **uniti questi sei grafi** per avere una visione complessiva di quelle che sono tutte le relazioni tra i nodi. Per fare ciò è stato costruito il seguente **multigrafo direzionato** *(che è stato poi usato per successive analisi)*:



Sono stati associati diversi colori in base alla relazione rappresentata.

**Statistiche di Rete**

In questa sezione verranno presentate una serie di **metriche** per descrivere le caratteristiche dei vari grafi presi in considerazione.

**Density:** indica **quanto è connessa la rete rispetto alla configurazione del grafo completo**: è il rapporto tra il numero di relazioni osservate nel grafo e il totale delle relazioni possibili.   
Nella tabella sotto sono riportate, *per i diversi grafi*, le **probabilità che dati due nodi i e j scelti casualmente, si osservi una connessione tra di loro**. In tutti i grafi questa **probabilità è molto bassa** *(intorno all’1%)*, mentre nel grafo completo, essendo una fusione di tutti gli altri grafi, si osserva un valore leggermente più alto.

Questi risultati sono in linea con le configurazioni dei grafi: sono abbastanza sparsi, presentano diversi nodi isolati e anche le componenti connesse che possono essere individuate sono comunque piccole.

|  | ***Family*** | ***Couples*** | ***Ex*** | ***Complete*** |
| --- | --- | --- | --- | --- |
| ***Densità*** | *1.4%* | *1.2%* | *0.9%* | *4%* |

**Reciprocity:** indica il **numero di relazioni bidirezionali rispetto al numero complessivo di relazioni nel grafo**.

Si riportano anche per questa le percentuali trovate per i quattro grafi considerati:

|  | ***Family*** | ***Couples*** | ***Ex*** | ***Complete*** |
| --- | --- | --- | --- | --- |
| ***Reciprocità*** | *0%* | *100%* | *77.42%* | *57.55%* |

Nel **grafo delle coppie** si hanno **solo relazioni bidirezionali** tra due nodi o nodi singoli, il che giustifica il valore del ***100%*** di reciprocità.

Nel caso della **famiglia** invece si ha la situazione opposta. Dato che il grafo ha una struttura simile a quella di un **albero genealogico** *(con relazioni di tipo figlio-padre)*, si ottiene una reciprocità nulla.

Meno drastica è la situazione ottenuta nel grafo degli ex, che differisce da quella delle coppie poiché non considera solo relazioni passate i rifiuti, rappresentati come relazioni non bidirezionali.

**Transitivity:** è una misura strettamente correlata al **concetto di triangoli**.

In generale, si considera il rapporto tra il **numero delle triple** *(sottografi con tre nodi)* ed il **numero dei triangoli** *(ovvero i sottografi in cui tutti i nodi sono in relazione tra di loro)*.

Per quanto concerne i grafi analizzati, sono stati ottenuti i seguenti risultati:

|  | ***Family*** | ***Couples*** | ***Ex*** | ***Complete*** |
| --- | --- | --- | --- | --- |
| **Transitività** | *0%* | *NaN* | *0%* | *21.71%* |

Analizzando i risultati ottenuti, si nota come **spesso il risultato derivi da una caratteristica intrinseca della relazione** analizzata.

Il grafo rappresentante le **relazioni familiari**, *ad esempio*, ha una struttura ad **albero genealogico**.

Dunque, **è impossibile ritrovare dei triangoli** in un grafo con tale struttura, poiché **un figlio dovrebbe essere il padre di suo nonno** *(creando un paradosso temporale in stile film sci-fi anni 80)*.

Per quanto riguarda il **grafo delle coppie**, si ottiene, come valore di transitività, ***NaN***. Ovvero, l’elaboratore non è in grado di calcolare il valore richiesto.  
Questo avviene poiché, *all’interno del grafo delle coppie*, le **componenti connesse hanno dimensione massima pari a due**, questo poiché le relazioni tra pinguini sono solo **monogame**. Dunque **non vi sono triple** e per calcolare la transitività sarebbe necessario effettuare una **divisione per zero**.

Per quanto riguarda il **grafo degli Ex**, strutturalmente **vi è la possibilità che vi sia un triangolo**, poiché, potenzialmente, tre pinguini potrebbero essere stati in relazioni tra loro, in periodi diversi. Tuttavia, è effettivamente una **situazione poco probabile** e che, *nel grafo analizzato*, non si verifica. Per questo motivo, la transitività relativa a questo grafo è pari a zero.

Nell’ultimo grafo considerato, il grafo completo, si ottiene invece un valore di transitività, ovviamente, più alto e diverso da zero, ad indicare che sono presenti un certo numero di triangoli nel grafo.

**Assortativity:** si tratta di una misura che ci permette di **osservare il grado di omofilia** all’interno del grafo. I valori che può assumere sono contenuti nell’***intervallo [-1;1]***. Il **valore 1** indica **perfetta omofilia** mentre il **valore -1** indica **perfetta eterofillia.** Nei grafi, i valori ottenuti sono i seguenti:

|  | ***Family*** | ***Couples*** | ***Ex*** | ***Complete*** |
| --- | --- | --- | --- | --- |
| **Assortativity** | *-0.008883013​* | *-1* | *-0.7659417* | *-0.516669* |

Dai risultati osservabili nella tabella, è possibile notare come all’interno del **grafo delle coppie** vi sia **perfetta eterofilia**. Questo implica che **non sono presenti relazioni di tipo omosessuali**. 

In secondo luogo, è possibile osservare come **nel grafo degli ex**,il valore dell’assortatività **non sia esattamente pari a -1**.Infatti, si nota che nel grafo è presente **una relazione di tipo omosessuale tra due pinguini.**

*E’ degno di nota anche il fatto che oltre ad essere dello stesso sesso, i due pinguini sono anche legati da un legame di parentela* *(essendo Kami il padre di Jouruiku).*

**Statistiche Nodali**

In questa sezione verranno presentate una serie di **misure di centralità** con l’obiettivo di capire se ci sono, all’interno dei grafi, dei **nodi più centrali** e se un nodo che è centrale rispetto a una certa misura lo è anche rispetto alle altre.

**Degree Centrality:** un nodo è centrale se è **connesso a molti altri nodi**. Maggiore è il valore della degree centrality associata ad un nodo, più quest’ultimo è centrale.

Nella tabella sottostante, per i diversi grafi, viene riportato il valore del nodo *(o dei nodi)* che presenta il **valore più alto di degree centrality**. Essendo i nostri grafi direzionati, nella tabella viene fatta la **distinzione tra in e out-degree centrality**.

| ***Max Degree Centr****.* | ***Family*** | ***Couples*** | ***Ex*** | ***Complete*** |
| --- | --- | --- | --- | --- |
| ***IN*** | *4* | *1* | *2* | *6* |
| ***OUT*** | *2* | *1* | *6* | *10* |

Nel **grafo delle coppie** il valore massimo, *sia per la in che per l’out degree centrality*, è 1, sensato dato che si osservano **solo relazioni monogame**.

Nel **grafo delle famiglie** il **massimo valore in out-degree è 2**, il che ha senso dato che ogni pinguino può avere al **massimo due genitori**.

Il **massimo valore per l’in-degree è invece 4**, ad indicare che i pinguini con il maggior numero di figli *(Man e Koko)* ne hanno quattro.

Nel **grafo degli ex** è interessante notare come **il massimo valore di out-degree** *(6)* è ottenuto da **Tera**, *una pinguina che ha rifiutato il maggior numero di spasimanti*.

Nel **grafo completo**, **Tera si conferma sempre il nodo con la out-degree centrality più alta**, mentre **per l’in-degree si distinguono Kuruma e Koko**, *quest’ultima* *probabilmente per la posizione di centralità rivestita nel grafo delle famiglie.*

**Closeness Centrality:** definisce **quanto un nodo è vicino agli altri nodi**.   
Si calcola effettuando la media tra la lunghezza dello shortest path dal nodo corrente a tutti gli altri nodi della rete.

I due tipi di closeness sono:

* **In-closeness centrality:** Misura quanto un nodo può essere facilmente raggiunto dagli altri nodi;
* **Out-closeness centrality:** Misura il modo in cui un nodo possa facilmente raggiungere gli altri nodi;

Il calcolo della closeness si può spiegare considerando il seguente sottoinsieme di nodi, preso dal Family Graph:



In-closeness(Kujou): = = 0.167 Nota: ***Nota:*** SHK e BHK hanno lunghezza 2

Out-closeness(Kujou): Nan (Nessun nodo uscente da Kujou)

**Betweenness Centrality:** è una misura che tratta il **concetto di centralità dei nodi in termine di quanti percorsi minimi attraversano uno specifico nodo** *(ovvero da quanti nodi deve essere attraversato un nodo per le comunicazioni)*.

All’interno dei grafi analizzati i **valori** trovati per questa metrica sono **piuttosto bassi**, come si può osservare nella seguente tabella:

|  | ***Family*** | ***Ex*** | ***Couples*** | ***Complete*** |
| --- | --- | --- | --- | --- |
| ***Min*** | *0.000e+00* | *0.000000​* | *0* | *0.000000​* |
| ***1st Qu*** | *0.000e+00​* | *0.000000​* | *0* | *0.002483* |
| ***Median*** | *0.000e+00​* | *0.000000​* | *0* | *0.011494* |
| ***Mean*** | *3.589e-05​* | *0.000241* | *0* | *0.040722​* |
| ***3rd Qu*** | *0.000e+00* | *0.000000​* | *0* | *0.056108* |
| ***Max*** | *1.210e-03* | *0.003327​* | *0* | *0.245034* |

Interessante è il valore ottenuto nel caso del **grafo delle coppie**, ma piuttosto ragionevole. Infatti, si ottengono **sempre valori pari a zero** poiché all’interno di tale grafo **sono presenti solo percorsi di lunghezza minore o uguale di uno**, dunque è impossibile che un nodo venga attraversato dal percorso minimo tra altri due nodi.

**Nel grafo degli ex e nel completo**, il pinguino con la maggiore Betweenness Centrality è **Tera**, già incontrato come nodo più centrale in precedenza.

Nel **grafo delle famiglie**, d’altra parte, il pinguino **più centrale è Hachi**.  
Questo è piuttosto sensato, poiché questa pinguina si trova al **centro di una delle famiglie più grandi** *(ovvero una delle componenti connesse più numerose)*.

**Eigenvector Centrality:** misura l’**importanza di un nodo** tenendo in considerazione l’**importanza dei suoi nodi vicini**.

|  | ***Family*** | ***Ex*** | ***Couples*** | ***Complete*** |
| --- | --- | --- | --- | --- |
| ***Min*** | *0.000e+00* | *0.000000​* | *0.000000​* | *0.003169​* |
| ***1st Qu*** | *0.000e+00​* | *0.000000​* | *0.000000​* | 0.027638 |
| ***Median*** | *0.000e+00​* | *0.000000​* | *0.000000​* | 0.060688​ |
| ***Mean*** | 0.04761 | 0.05412​ | 0.04468 | 0.092582 |
| ***3rd Qu*** | *0.000e+00* | *0.000000​* | 0.04754​ | 0.124798 ​ |
| ***Max*** | 0.51642 | *0.54535​* | 0.31252 | 0.460284 |

I valori riportati **non appaiono particolarmente significativi** e ciò deriva dalla sparsità dei grafi presi in considerazione.   
I **nodi con eigenvector centrality maggiore** sono:

* **Koko***(nel grafo delle famiglie)*;
* **Tera** *(nel grafo degli ex)*;
* **Mikage** *(nel grafo delle coppie)*;
* **Nishi** *(nel grafo completo).*

**Graph Models**

Per i graph models sono stati applicati i modelli visti in classe ai grafi scelti *(Family, Ex, Couples, Complete)*.

**Modelli semplici**

Come primo modello è stato considerato il **Binomial Random Graph Omogeneo** *(mod0)*, usando solo il coefficiente *edges* che rappresenta il numero di archi nella rete.

Da questo abbiamo ottenuto un **coefficiente negativo e significativo**. Questo significa che nel grafo ci sono meno legami del previsto *(di quanti ce ne saremmo aspettati)*.

Per il **Binomial Random Graph Non Omogeneo** *(mod1)* invece sono stati aggiunti i coefficienti ***sender*** e ***receiver***, che si riferiscono alle relazioni nel grafo diretto *(si basano sulle dummy, quindi ci saranno un sender e un receiver principali coi quali verranno confrontati gli altri 58 nodi del grafo)*.

***Sender*** rappresenta l’**insieme degli archi uscenti** da un nodo. Se l’**effetto dovesse essere negativo**, il sender avrebbe un numero di archi in uscita inferiore rispetto al sender principale che si sta confrontando *(si comporta come una variabile dummy)*.

***Receiver*** rappresenta il medesimo concetto ma per **gli archi entranti nei nodi**.

I coefficienti trovati, in entrambi i casi, hanno assunto valori ***-Inf*** o **non significativi**, quindi si è deciso di **non includerli nei modelli successivi**.

Per la fase di **model selection** sono state usate le metriche **AIC e BIC**.

|  | ***Family*** | | ***Ex*** | | ***Couples*** | | ***Complete*** | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | *AIC* | *BIC* | *AIC* | *BIC* | *AIC* | *BIC* | *AIC* | *BIC* |
| *mod0* | *507* | *513* | *355* | *361* | *455* | *461* | *1152* | *1158* |
| *mod1* | *451* | *705* | *325* | *547* | *561* | *1013* | *1268* | *1954* |

*Tra i due, il modello che è stato mantenuto per confronti con i modelli successivi è il mod0.*

**Modello p1**

Dopo aver accertato la **non significatività** dei coefficienti associati a ***senders*** e ***receivers***, per la costruzione del **modello p1** *(mod 3)* verranno inseriti solo gli **effetti degli archi e di mutual**, *il parametro associato alla reciprocità delle relazioni*.

Per il **grafo delle coppie**, **non è possibile terminare l’esecuzione del modello**, *forse a causa dei troppi nodi isolati*.

Nel **grafo delle famiglie**, *come atteso*, il **coefficiente associato a mutual è -Inf**, dato che la relazione *figlio-genitore* è *“gerarchica”* e **non può essere mutuale**.

Nel **grafo degli ex** e nel **grafo completo** il coefficiente associato a **mutual** è sempre **significativo e positivo**.

Per questi due grafi, sia guardando l’AIC che il BIC, il **modello p1 viene sempre preferito al modello nullo**, *con il solo effetto degli archi*.

| ***​​*** | ***Family​​*** | | ***Ex​​*** | | ***Complete​​*** | |
| --- | --- | --- | --- | --- | --- | --- |
| *​​* | ***mod0****​* | ***mod3*** | ***mod0*** | ***mod3*** | ***mod0*** | ***mod3*** |
| ***AIC*** *​* | *507* | *NaN* | *355* | *248* | *1152* | *960* |
| ***BIC*** *​* | *513​* | *NaN* | *361* | *260* | *1158* | *973* |

**Modelli con attributi nodali**

L’**attributo nodale** presente nel dataset, ed utilizzato per la costruzione dei modelli, è il **genere** dei pinguini.

Sono stati considerati sia il **main effect** che l’effetto relativo all’**omofilia**.

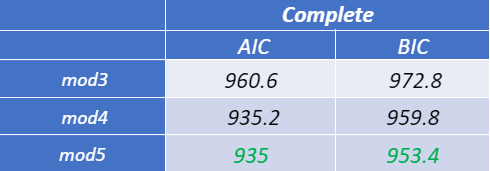
In questo modo, è stato costruito il modello indicato come *“mod4”*, eseguito per tutti e quattro i grafi. Il modello costruito assume, dunque, la seguente struttura:

*mod4: mod3 + node\_factor + node\_match;​*

Nella maggior parte dei casi, i **coefficienti** relativi agli effetti dell’attributo sono risultati **non significativi**.  
Nel caso del grafo relativo alle **coppie**, i coefficienti assumono valore ***NaN e -Inf***.   
Ciò accade poiché le **relazioni amorose**, *attualmente in corso*, tra i pinguini dell’acquario di Kyoto sono **solo eterosessuali**.

| ***​​*** | ***Family​​*** | | ***Ex​​*** | |
| --- | --- | --- | --- | --- |
| *​​* | ***mod0****​* | ***mod4****​* | ***mod3****​* | ***mod4****​* |
| ***AIC*** *​* | *510​* | *507​* | *247.7​* | *237.6​* |
| ***BIC*** *​* | *513​* | *528.4​* | *260​* | *262.1​​* |

Differentemente dagli altri, il **grafo completo** ha l’attributo di **omofilia significativo**, e dunque è stato possibile **costruire anche un quinto modello**, *solo per quest’ultimo grafo*:



*mod5: edges + mutual + node\_match;*

Modello che, anche se di poco, si comporta meglio rispetto ai modelli precedentemente utilizzati per il grafo completo.

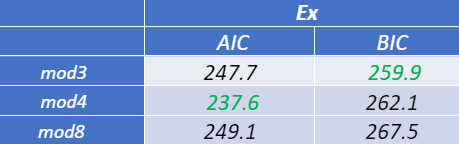
**Social Circuit  
f**Per ogni modello sono stati selezionati **selezionato soltanto gli attributi significativi** ed è stato aggiunto il **coefficiente *gwdsp*** per gli **alterneting k-2 paths**. Quindi abbiamo considerato il seguente modello:

*mod 8:* *edges + mutual + node\_match + gwdsp*

Si può notare che il **coefficiente *gwdsp*** risulta **significativo e negativo** nel **grafo completo** e nel **grafo della famiglia**.   
Dal punto di vista interpretativo, ciò implica che risulta una **minor tendenza alla formazione di cluster** rispetto a quelli che si otterrebbero in **maniera casuale**.

D’altra parte, nel caso dei **grafi degli ex e delle coppie**, il **coefficiente *gwdsp*** risulta **non significativo.**   
Di seguito riportiamo il confronto tra il modello mod8 ed i modelli analizzati in precedenza:

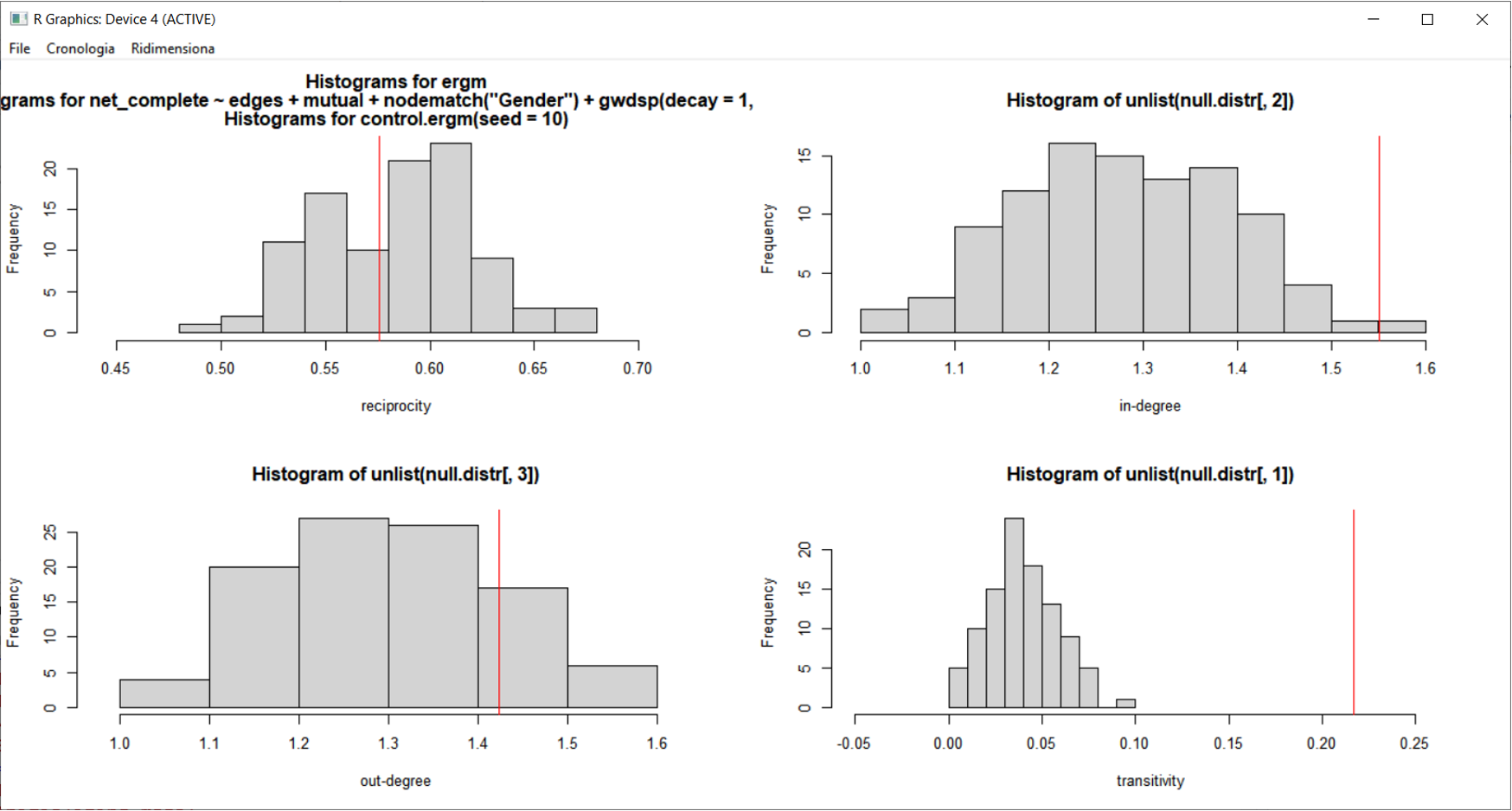
| ***​​*** | ***Family​​*** | | ***Couples*** | | ***Complete​​*** | |
| --- | --- | --- | --- | --- | --- | --- |
| *​​* | ***mod0****​* | ***mod8*** | ***mod0*** | ***mod8*** | ***mod5*** | ***mod8*** |
| ***AIC*** *​* | *507* | *474.7* | *455.1* | *410.8* | *935* | *926* |
| ***BIC*** *​* | *513​* | *487* | *461.2* | *423.1* | *953.4* | *951* |

In verde vengono riportati i modelli che hanno dimostrato valori minori di AIC e BIC e che abbiamo infine selezionato. Questi modelli verranno adesso posti ad una fase di diagnostica.

**Diagnostica sui modelli**

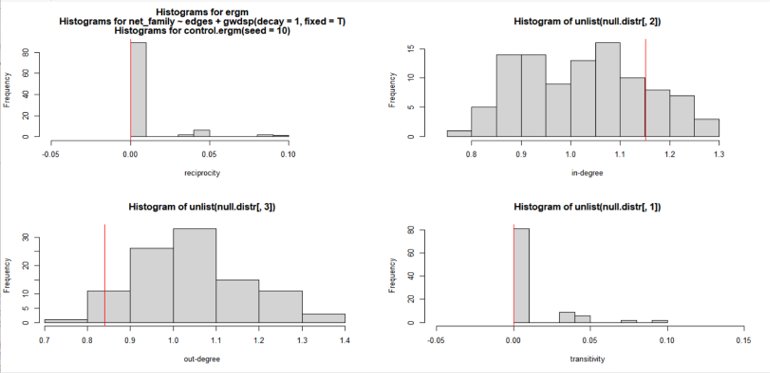
Per capire se il modello scelto *(con AIC e BIC minori)*, riesca effettivamente a **cogliere le caratteristiche del grafo osservato**, si **confrontano** il valore della **statistica osservata** sul grafo *(la linea rossa nella figura sotto)* con la **distribuzione ottenuta applicando il modello** scelto a delle simulazioni di grafi con metodo Monte-Carlo.

***Diagnostica Modello Social Circuit - Grafo completo:  
f***

Dalle distribuzioni a fianco si può osservare come il modello scelto, *in questo caso il Social Circuit*, riesca a cogliere **bene la reciprocità** e **abbastanza bene l’out-degree**, mentre **non riesce a predire l’in-degree** e la **transitività.**

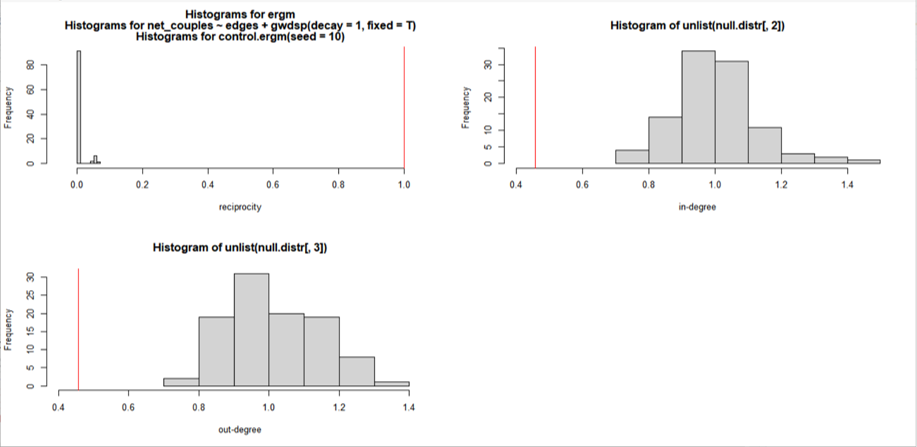
Ripetendo la diagnostica sul modello 5 *(con AIC e BIC solo leggermente superiori)*, si ottiene una previsione migliore sull’in-degree, ma non si riesce comunque a cogliere la transitività.

***Diagnostica Modello Social Circuit - Grafo della famiglia:***

Anche in questo caso le distribuzioni in figura **non presentano dei risultati completamente soddisfacenti**. 

Infatti, se il Social Circuit Model riesce a rappresentare, più o meno, **bene la reciprocità e la transitività**, *in quanto ponendosi abbastanza vicino al picco della distribuzione*, d’altra parte, esso non riesce nell’intento di catturare **l’out-degree e l’in-degree**

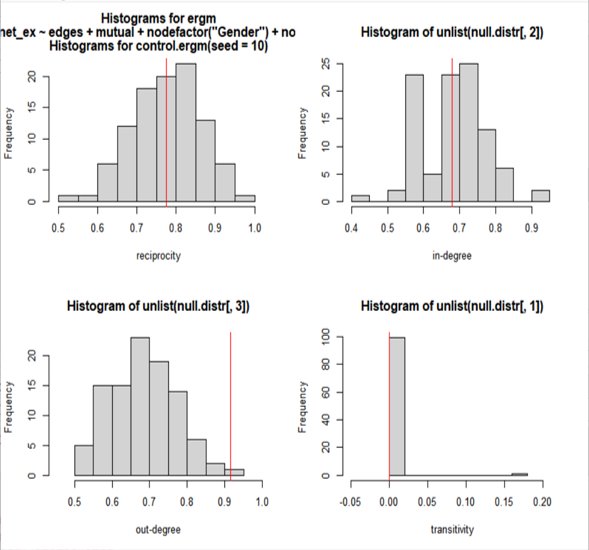
***Diagnostica Modello Social Circuit - Grafo delle coppie***

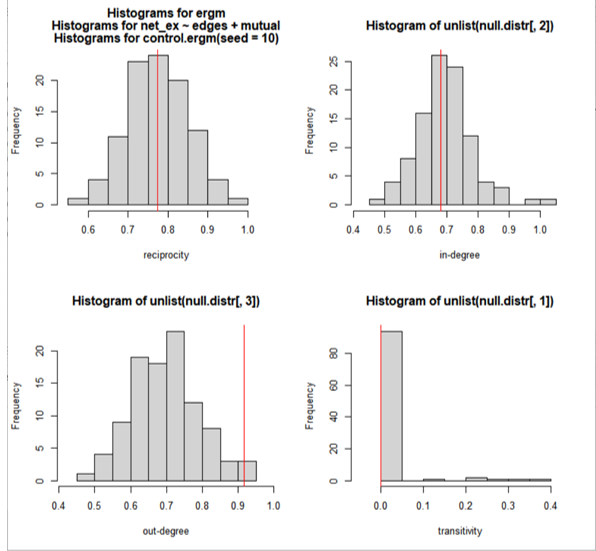
Rispetto ai due casi precedenti qui la situazione è pessima. Infatti, **la statistica osservata nel grafo considerato si distacca tantissimo dalla distribuzione simulata**. Ciò è visibile dai tre istogrammi ed è dovuto alla struttura inusuale del grafo delle coppie.

Inoltre in questo caso **non è presente l’istogramma relativo alla transitività** perché, come visto in precedenza, nel grafo delle coppie non è possibile calcolare la transitività *(dato che vi sono solo relazioni tra coppie di nodi e non ci sono triangoli)*.

***Diagnostica Modelli P1 e con attributi nodali - Grafo degli ex***

Per quanto concerne il **grafo degli ex**, sono stati selezionati **due modelli**, poiché AIC e BIC, nell’ultima esecuzione, indicavano due diversi modelli come migliori.

Il BIC predilige il **modello P1**, l’AIC, d’altra parte, predilige il **modello con attributi nodali**.



In generale, entrambi i modelli sono piuttosto **accurati**, eccetto che per l’out-degree in cui i modelli sono tendenzialmente meno accurati.

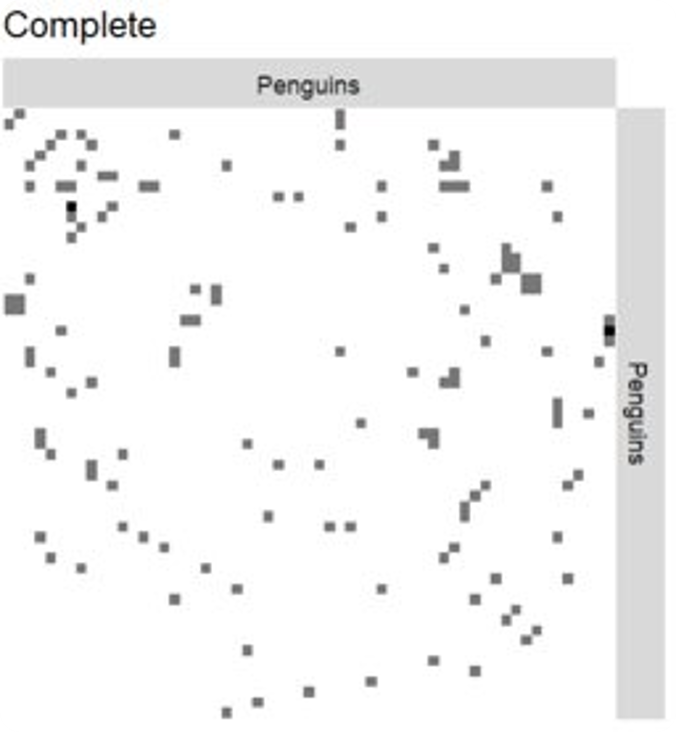
**Recap dei modelli**

Nella tabella sottostante sono riportati i vari **step effettuati nella definizione dei modelli**.   
A partire dai modelli più semplici, si aggiungono parametri per catturare diversi effetti e via via rimuoviamo i coefficienti non significativi o con cui il modello non riesce a girare.

Sono stati riportati anche i **modelli** che si è tentato di eseguire, ma per cui **non è stato possibile completarne l’esecuzione**, *per diversi motivi*.

| Models (seed = 10)​ | Description​ | Features​ |
| --- | --- | --- |
| *Model 0​* | *Null model BRG​* | *edges​* |
| *Model 1​* | *NH-BRG​* | *Model 0 + sender + receiver* |
| *Model 2​ (it doesn't run)​* | *NH-BRG​* | *Model 1 + mutual* |
| *Model 3​ (not run with couples)​* | *NH-BRG​* | *Model 0 + mutual* |
| *Model 4​* | *NH-BRG with nodal attributes (main + homophily effect)​* | *Model 3 + node\_factor + node\_match​* |
| *Model 5​* | *NH-BRG with nodal attributes (only homophily effect)​* | *Model 4 - node\_factor* |
| *Model 6​ (degeneracy)​* | *Markov Graph Model​* | *Model 5 + triangle term, in and out stars of order two* |
| *Model 7​ (it doesn't run)​* | *Social Circuit Model with alternating k-triangle and alternating k-2-path​* | *Model 3 + node\_match + gwdsp + gwesp​* |
| ***Model 8​*** | *Social Circuit Model with alternating k-2-path​* | *Model 3 + node\_match + gwdsp* |

**Block Models**

L’obiettivo di questa sezione è capire se si possono osservare **clusters di nodi**, all’interno dei grafi analizzati, che abbiano caratteristiche analoghe.

Con la libreria sbm di R, applicando il modello Bernoulliano o di Poisson possono essere fatte delle stime per capire se si evidenziano tali tendenze nei grafi.

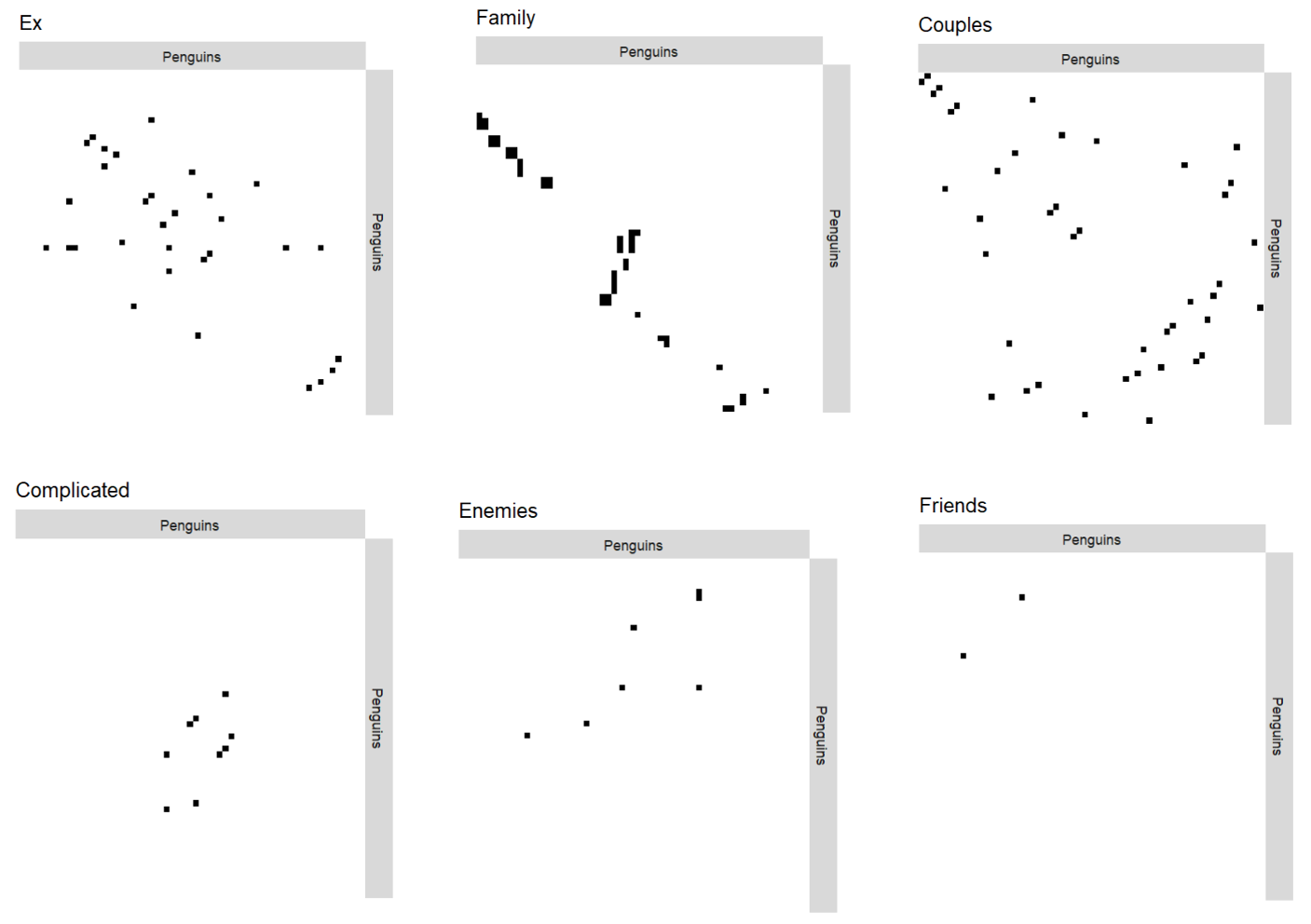
A prescindere dal grafo analizzato, guardando sia all’**ICL** che alla **loglikelihood**, il **modello selezionato** risulta essere sempre quello con un **unico cluster**.

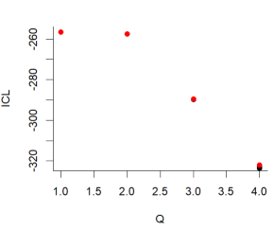
*Come si nota dalla rappresentazione grafica*, le **matrici di adiacenza sono tutte molto sparse**.

Questo potrebbe spiegare il motivo per cui è difficile cercare un numero di cluster maggiore di uno.

Il grafo che risulta **più interessante** è quello rappresentante le **relazioni familiari**, in cui **i punti sembrano sovrapporsi**, *formando dei blocchi*.

Questo è spiegato dal fatto che si possono individuare famiglie più numerose di altre, ovvero con più collegamenti.



D’altra parte, per tentare di effettuare un’**analisi più omnicomprensiva**, si è deciso di proseguire con l’analisi dei modelli, senza fermarsi al primo modello selezionato tramite ICL.

Si è cercato di fare un **trade off tra il valore dell’ICL ottenuto ed il numero di cluster presenti dal modello**. Nella maggior parte dei casi, **le differenze in termini di ICL appaiono piuttosto significative**. Invece, **per il grafo della famiglia**, la **differenza** tra il modello con un solo cluster e con due cluster risulta essere **molto meno marcata**, come visibile nel grafico.

Solo per il **grafo delle coppie non è stato scelto alcun modello**, poiché, avendoli provati tutti, si è verificato che, in ogni modello, **tutte le istanze venivano raggruppate in un singolo cluster**, lasciando gli altri vuoti.

I modelli scelti, per grafo, sono stati i seguenti:

* ***Family → 2 Cluster***;
* ***Ex → 3 Cluster***;
* ***Complete → 2 Cluster***.

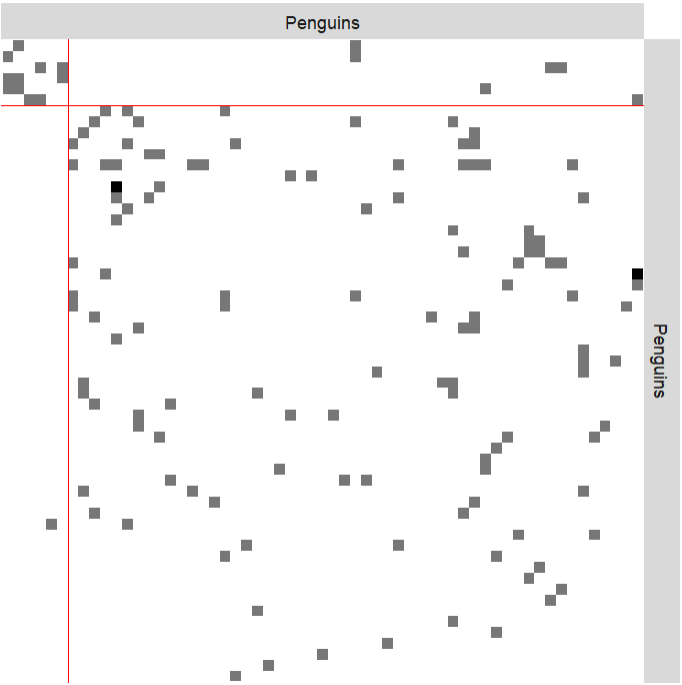
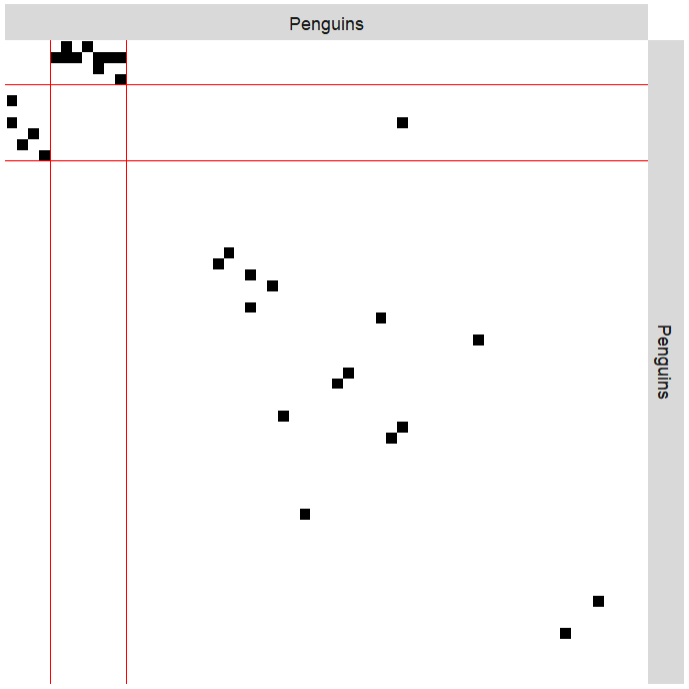
Il **numero di osservazioni per cluster** è riportato nella seguente tabella:

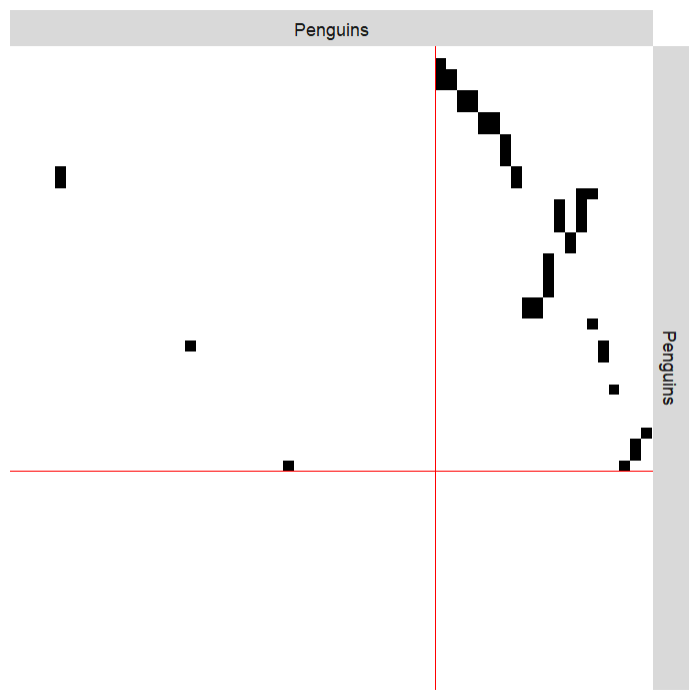
|  | ***Family*** | | ***Ex*** | | | ***Complete*** | |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *Cluster\_id* | *1* | *2* | *1* | *2* | *3* | *1* | *2* |
| *n\_obs* | *39* | *20* | *4* | *6* | *49* | *6* | *53* |

In generale, si ottengono sempre **cluster molto sbilanciati** e che portano a pensare che il modello con un **singolo cluster sia effettivamente quello da preferire**.

Diverso è il caso del **grafo della famiglia**, in cui si ottiene una **clusterizzazione che è piuttosto ragionevole**, come atteso dalle precedenti valutazioni.

In basso, vengono riportati i **plot relativi alla clusterizzazioni appena analizzate**:

***Family*** ***Ex*** ***Complete***



**Conclusioni**

Alla luce dei risultati ottenuti nelle fasi precedenti, si è giunti alle seguenti considerazioni:

* I criteri AIC e BIC portano a selezionare il **Social Circuit with edges and alternating k-2 paths** per quanto riguarda i grafi ***completo, famiglia e coppie***. Nonostante ciò, la **fase diagnostica** ha dimostrato come questa scelta non sia completamente appropriata.
* Per quanto riguarda il **grafo degli ex**, è stato selezionato il modello **P1 with nodal attributes**, la cui fase di diagnostica ha confermato la correttezza della scelta.
* Le **statistiche nodali** di *betweenness centrality* e *eigenvector centrality* ci indicano il pinguino ***Tera***come il nodo più centrale sia nel grafo degli Ex che nel grafo completo.
* Il pinguino ***Koko*** risulta essere il nodo con il valore più alto di *in-degree* e di *eigenvector centrality* all’interno del grafo della famiglia ed il massimo valore di *out-degree* nel grafo completo

Si noti infine, come i pinguini sopra elencati siano classificati come *“devilsh”* all’interno dei **metadati** allegati al dataset analizzato.  
Questa classificazione dà forza ai risultati osservati; infatti è ragionevole supporre che **pinguini dispettosi siano più attivi socialmente**, in quanto essi sono portati a relazionarsi di più.