# Gravitational Algorithm for Anomaly Detection

Javier Fumanal Idocin[1], Alfonso Induráin[1], María Minárová[2], Mikel Ferrero Jaurrieta[1], Humberto Bustince[1]

3rd February, 2022

[1] Public University of Navarre
[2] Slovak University of Technology in Bratislava

*16th International Conference on Fuzzy Set Theory and Applications*
Liptovský Ján, Slovak Republic

Universidad Pública de Navarra
Nafarroako Unibertsitate Publikoa

# Table of Contents

# Gravitational clustering

# Classical gravitational clustering

What is cluster analysis?

- maximal mutual similarity inside a cluster
- minimal similarity between clusters

**Gravitational clustering algorithm (Wright, 1977)**
alternating of two steps:

(A) **Attracting** - driven by Newton gravitational law
controlling parameter: $\delta$ - maximal shift (of the fastest particle) within one time step

(C) **Clustering**
controlling parameter: $\epsilon$ - limiting distance of particles

- Setting:
  - initial time to zero
  - influencing parameters $\epsilon$ and $\delta, \delta < \epsilon$
  - initial mass of each particle to 1

# Algorithm

- Repeating (i)-(vi) until 1 particle remains
  (i) checking the distances of particles. If less than $\epsilon$, unification in new one: mass = sum of masses, position in the center of gravity
  (ii) movement driving function $\forall p_i$, $\forall [t, t+dt]$

$$\vec{g}_i(t) = \frac{1}{2} G \sum_{j \neq i} \frac{1}{m_i(t)} \frac{\vec{s}_j(t) - \vec{s}_i(t)}{|\vec{s}_j(t) - \vec{s}_i(t)|^3} dt^2, \quad G > 0$$

  (iii) the fastest particle: $I = \arg(\max_i\{|\vec{g}_i|(t)\})$, $p_I$
  (iv) current time step length computation: $|\vec{g}_I(t)| = \delta \Rightarrow dt(t)$
  (v) new position of each particle $p_i$: $\vec{s}_i(t + dt(t)) = \vec{s}_i(t) + \vec{g}_i(t)$
  (vi) $t \leftarrow t + dt(t)$

# Anomaly detection and Gravitational Clustering

**Borgia Clustering.** Generalizing the gravitational clustering algorithm to form communities like the conquests of Cesare Borgia.


**(a)**

**(b)**

**Figure 1:** **Italy in times of the Borgias.** Italy's lordships before Cesare Borgia's campaign as Commander in Chief of the Papal Army. Each colour represents a different faction.

---

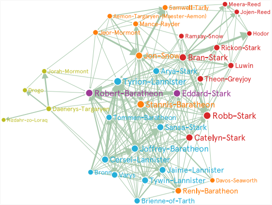
Fumanal-Idocin, J., Alonso-Betanzos, A., Cordón, O., Bustince, H., & Minárová, M. (2020). Community detection and social network analysis based on the Italian wars of the 15th century. *Future Generation Computer Systems*, 113, 25-40.

**Borgia Clustering.**

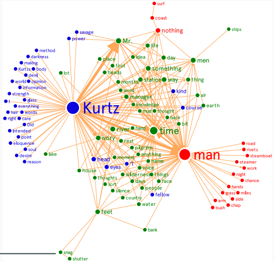
1. Different natural affinities between parties $\rightarrow$ affinity functions.
2. Balance of power $\rightarrow$ particle mass.
3. Scaling size: small communities differ from big communities $\rightarrow$ change the behaviour while communities grow.

Fumanal-Idocin, J., Alonso-Betanzos, A., Cordón, O., Bustince, H., & Minárová, M. (2020). Community detection and social network analysis based on the Italian wars of the 15th century. *Future Generation Computer Systems*, 113, 25-40.

# Borgia Clustering



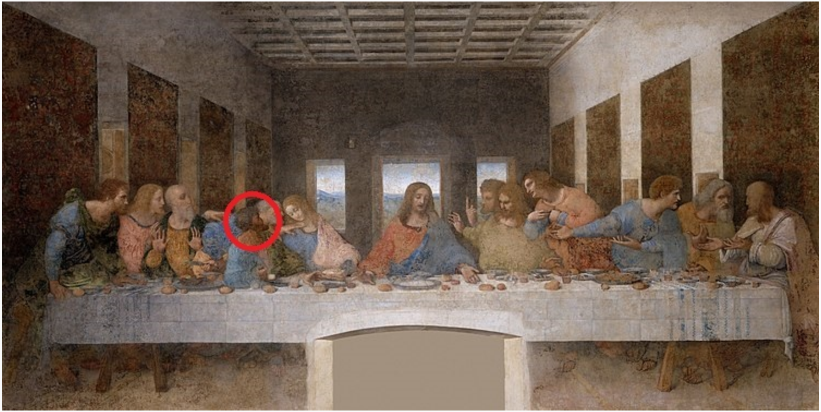





Fumanal-Idocin, J., Alonso-Betanzos, A., Cordón, O., Bustince, H., & Minárová, M. (2020). Community detection and social network analysis based on the Italian wars of the 15th century. *Future Generation Computer Systems*, 113, 25-40.

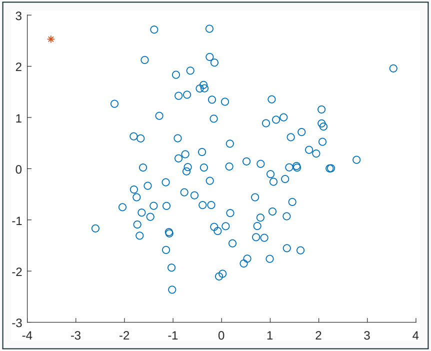
# There is always the weird one...

# There is always the weird one...



# A challenge... and an opportunity

Sometimes particles are **too far** away... bad results!



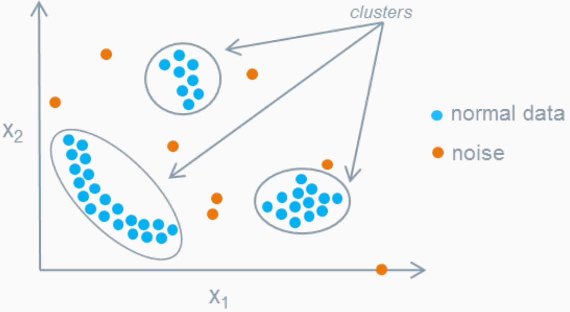
# A challenge... and an opportunity

Can we exploit this?

# Modifications to the original Gravitational Clustering

# Anomaly detection

**Anomaly**. A rare item or observation which deviate significantly from the majority of the data (Hawkins, 1980).



# Anomaly clusters

If a particle/cluster is "too far away" is an anomaly!

Two checks:

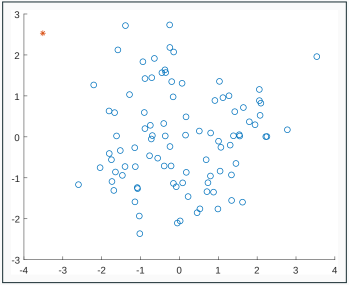
1. How **far**: percentage of the simulation that took place before it joined the rest of the observations
2. How **big**: if a lot of observations are close, then they are not anomalies. (Cluster size)

# Anomaly clusters

**Setting**. $P =$ Maximum anomaly cluster size. $T =$ coupling time. If a particle/cluster takes up more than $T$ to join other cluster can be considered an anomaly.

# Distances and particles

- Anomalies can affect the intuitive process of aggregation forces: a distant particle can create incorrect clustering configurations.



# Distances and particles

- Solution: use a more general expression for the aggregation of forces to mitigate extreme values.

# Generalizing the distance aggregation of forces

From

$$g_i^d(t) = \sum_{j \neq i} \frac{1}{m_i(t)} \frac{s_j^d(t) - s_i^d(t)}{|\vec{s}_j(t) - \vec{s}_i(t)|^3} dt^2(t)$$

to

$$g_i^d(t) = F_{j \neq i} \frac{1}{m_i(t)} \frac{s_j^d(t) - s_i^d(t)}{|\vec{s}_j(t) - \vec{s}_i(t)|^3} dt^2(t)$$

# FG-Sugeno integral

The discrete Sugeno integral with respect to the fuzzy measure $m$ is defined as a function $S_m : [0,1]^n \to [0,1]$, given for every $\mathbf{x} = (x_1, \ldots, x_n)$ (Bardozzo et al., 2020)

$$S_m(\mathbf{x}) = F\{G(x_{\sigma(i)}, m(A_i))|i = 1, .., n\} \tag{1}$$

where $\mathbf{x}_\sigma$ is an increasing permutation of $\mathbf{x}$ such that $0 \leq x_{\sigma(1)} \leq \cdots \leq x_{\sigma(n)}$. With the convention that $x_{\sigma(0)} = 0$, $A_i = \{(i), (i+1), \ldots, (n)\}$, $F$ is a n-ary function and $G$ is a bi-variant function.

For our experimentation, we take as $F$ the summation and $G$ as the product.

# Ad-hoc solution for the Sugeno integral

- We are aggregating vectors of forces (not numbers).

1. We sort the vector forces based on their moduli.
2. We multiply each vector to the correspondent measure.
3. We sum the resulting vectors.

# Experimental results

# Datasets utilized

**Table 1:** Imabalance ratio, features and samples for each dataset studied.

| Dataset | Imb. Ratio | Features | Samples |
|---|---|---|---|
| Ecoli1 | 3.36 | 7 | 336 |
| Ecoli3 | 8.6 | 7 | 336 |
| Glass6 | 6.38 | 9 | 214 |
| Wisconsin | 1.86 | 8 | 683 |
| Yeast3 | 8.1 | 8 | 1484 |

Keel repository: `https://sci2s.ugr.es/keel/imbalanced.php`

# Results

**Target metric.** F1 score

**Table 2:** F1 score for the proposed algorithm using different aggregation functions for the attraction forces.

| Dataset | Aggregation | |
|---|---|---|
| | Sum | Ad-hoc FS-Sugeno like |
| Ecoli1 | **0.8468** | 0.8350 |
| Ecoli3 | 0.9285 | **0.9432** |
| Glass6 | **0.9254** | 0.9057 |
| Wisconsin | **0.8062** | 0.7954 |
| Yeast3 | **0.9324** | **0.9324** |

# Decision making

- Different aggregations $\rightarrow$ different solutions to combine.
- In order to add more diversity we can use some well-known tactics:
  1. Sampling only a % of the observations.
  2. Using only some features.
- We generate $k$ different models using randomly chosen aggregation, features and samples.
- Majority vote as final decision.

# Results for decision making

**Table 3:** F1 score for the anomaly-Grav algorithm using different numbers of models.

| Datasets | Number of Models | | | |
|---|---|---|---|---|
| | 5 | 10 | 15 | 20 |
| Ecoli1 | 0.8841 | 0.8900 | 0.8947 | **0.8952** |
| Ecoli3 | 0.8887 | 0.8918 | 0.8943 | **0.8949** |
| Glass6 | 0.8925 | 0.8925 | 0.8938 | **0.8944** |
| Wisconsin | 0.8767 | 0.8904 | 0.8956 | **0.8968** |
| Yeast3 | 0.8884 | 0.8897 | 0.8916 | **0.8925** |

# Conclusions and Future lines

# Conclusions

- We proposed an adaptation of the gravitational clustering to perform anomaly detection.
- Performance measured using F1 score is good!
- Different aggregations than the sum can lead to better results.
- Decision making schemes seems to be very successful.

# Future Lines

- Use more aggregation functions.
- Develop further the decision making scheme.

*Thanks for your attention*

Any questions?

javier.fumanal@unavarra.es