Digital Media and Social Networks

Kleomenis Katevas
k.katevas@qmul.ac.uk
https://minoskt.github.io

Lecture 4: Network Analysis

Queen Mary
University of London

In this Lecture

- In this lecture we introduce the concept of network analysis for different social networks.
  - We will look at more natural social networks.
  - But we will still use standard measures and methods.

OSNs in early days (Mislove et al IMC'07 paper)

- Presents large-scale measurement study and analysis of the structure of multiple online social networks
  - 11 M users, 328 M links

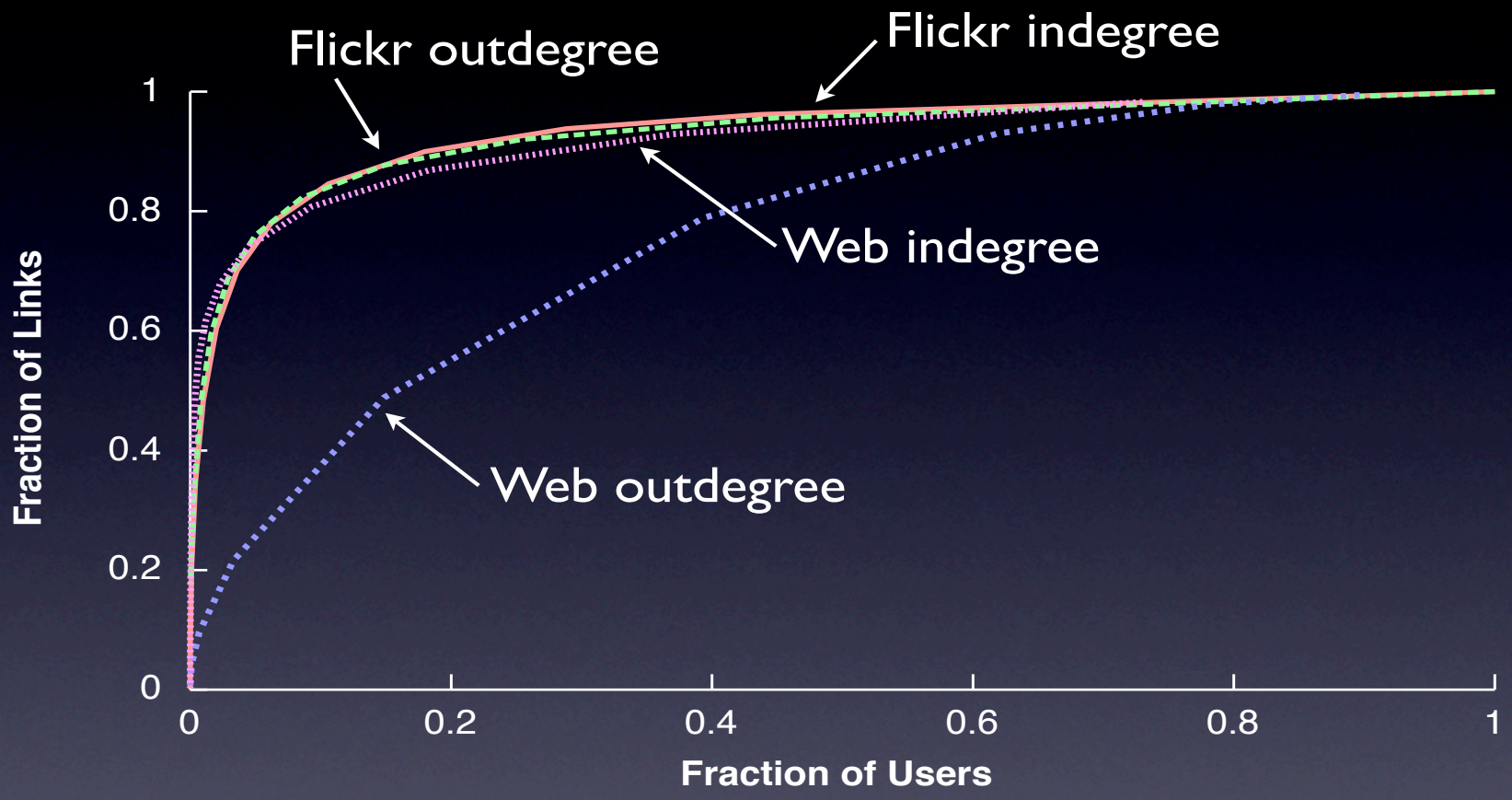
- Data from four diverse online social networks
  - Flickr: photo sharing
  - LiveJournal: blogging site
  - Orkut: social networking site
  - YouTube: video sharing

- Our goals are two-fold:
  - Measure online social networks at scale
  - Understand static structural properties

High-level data characteristics

	Flickr	LiveJournal	Orkut	YouTube
Number of Users	1.8 M	5.2 M	3.0 M	1.1 M
Avg. Friends per User	12.2	16.9	106.1	4.2

How are the links distributed?



- Distribution of indegree and outdegree is similar
  - Underlying cause is *link symmetry*

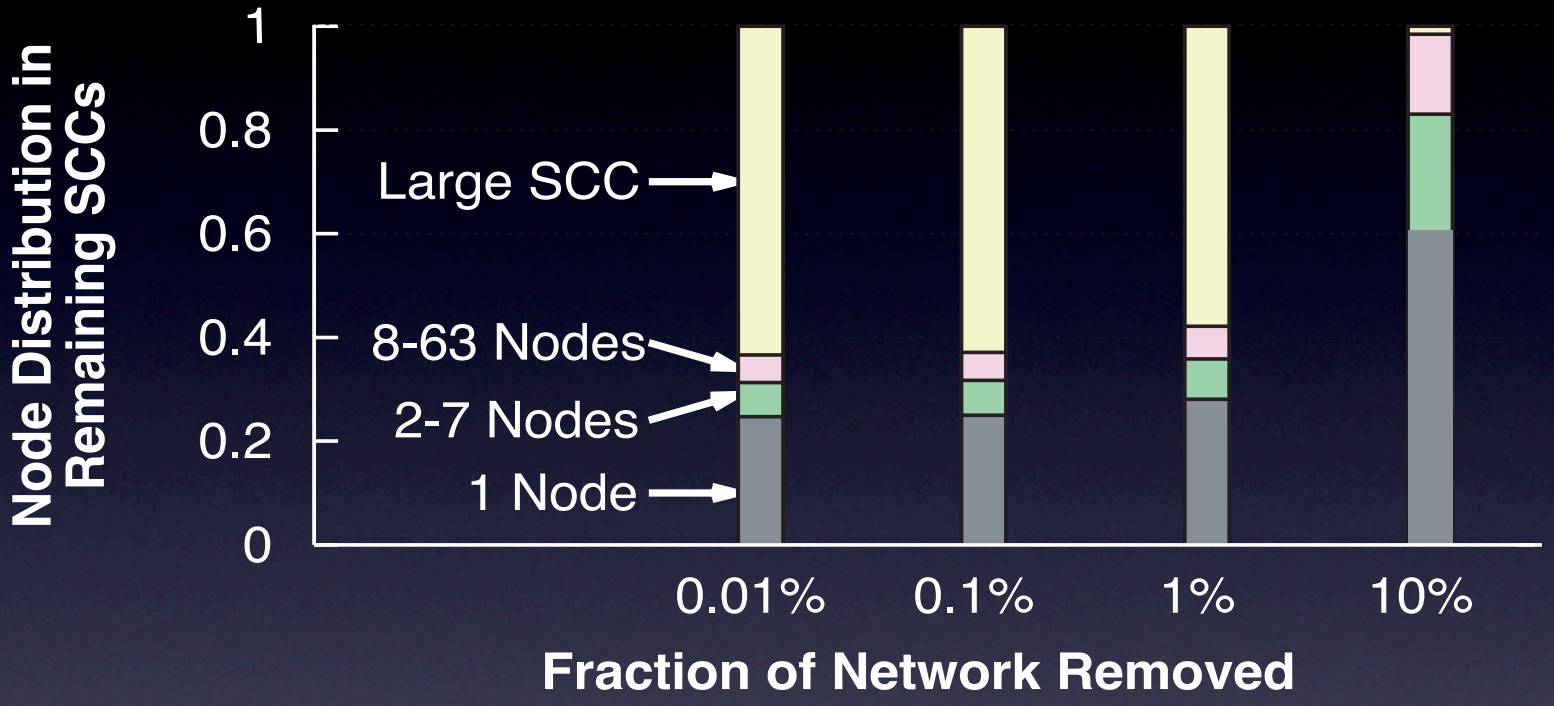
Link symmetry

- Social networks show high level of link symmetry
  - Links in most networks are directed

	Flickr	LiveJournal	Orkut	YouTube
Symmetric Links	62%	73%	100%	79%

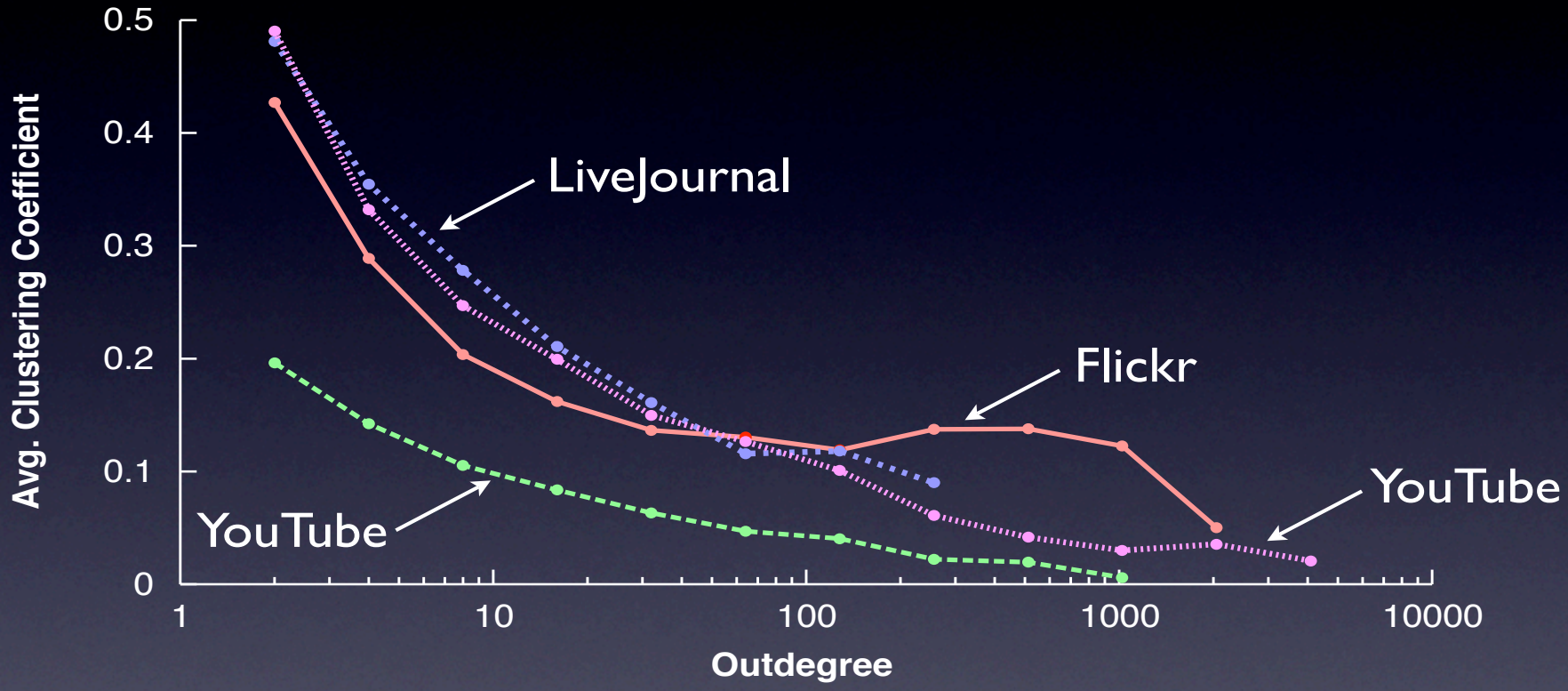
- High symmetry increases network connectivity
  - Reduces network diameter

Does a core exist?



- Yes, networks contain core consisting of 1-10% of nodes
  - Removing core disconnects other nodes

Are the fringes more clustered?



- Low-degree users show high degree of clustering
  - Networks are small-world, may be scale-free

Implications of network structure

- Network contains dense core of users
  - Core necessary for connectivity of 90% of users
  - Most short paths pass through core
  - Could be used for quickly disseminating information

- Fringe is highly clustered
  - Users with few friends form mini-cliques
  - Similar to previously observed offline behavior
  - Could be leveraged for sharing information of local interest

User Generated Content on YouTube

Name	Category	# Videos	Tot. views	Tot. length	Data collection period
YouTube	Ent	1,687,506	3,708,600,000	15.2 years	Dec 28, 2006 (crawled once)
YouTube	Sci	252,255	539,868,316	1.8 years	Jan 14 - 19, '07 (daily), Feb 14, '07, Mar 15, '07 (once)
Daum	All	196,037	207,555,622	1.0 year	Mar 1, 2007 (crawled once)

The graph shows that 10% of the top popular videos account for nearly 80% of views, while the rest 90% of the videos account for very requests.

Pareto Principle (or 80-20 rule) ?

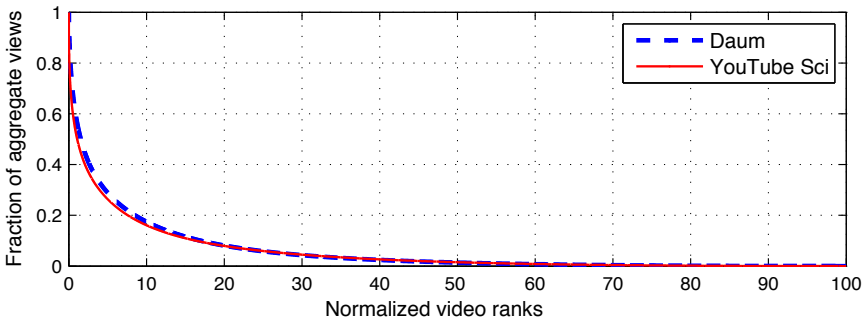
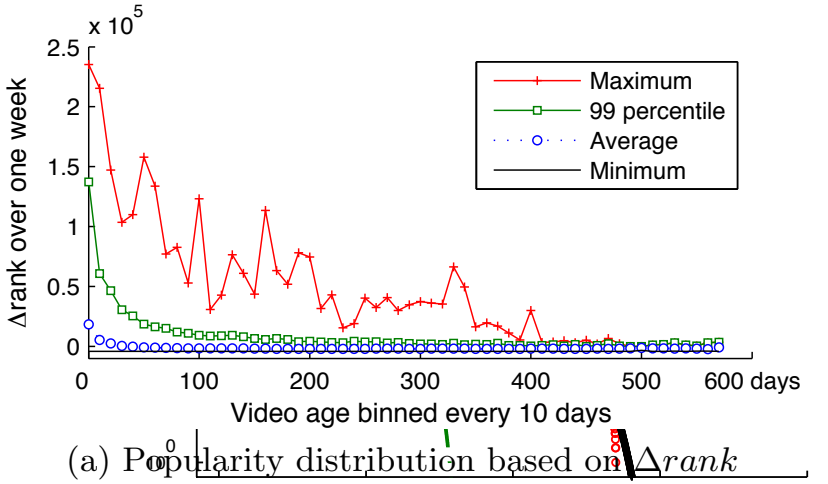
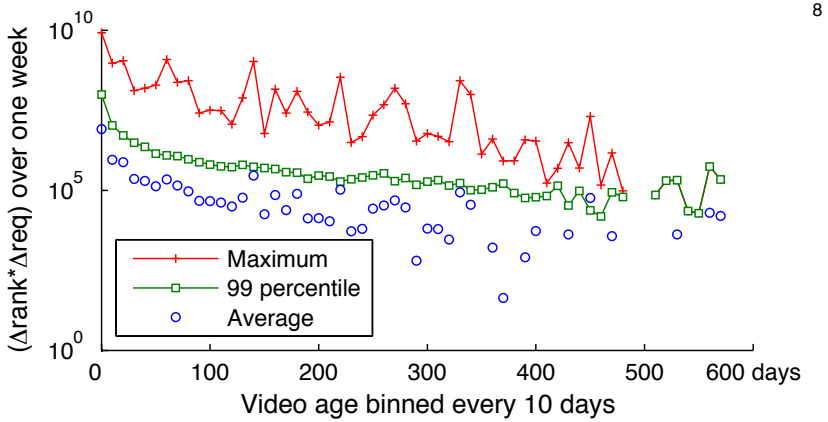


Figure 2: Skewness of user interests across videos

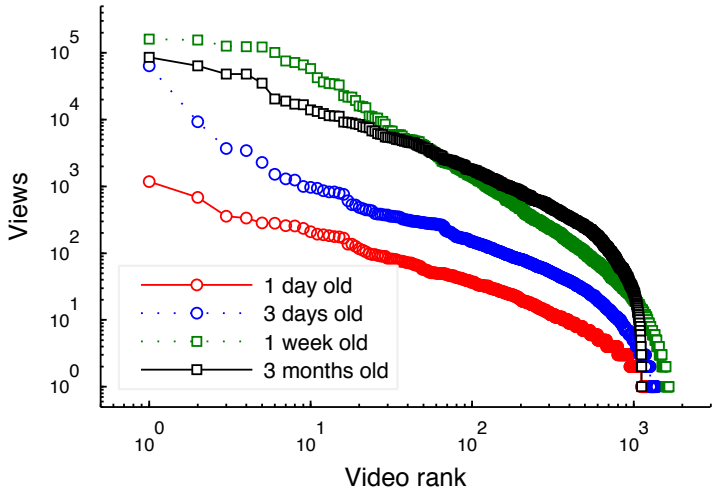
More video: age!



(a) Popularity distribution based on $\Delta rank$



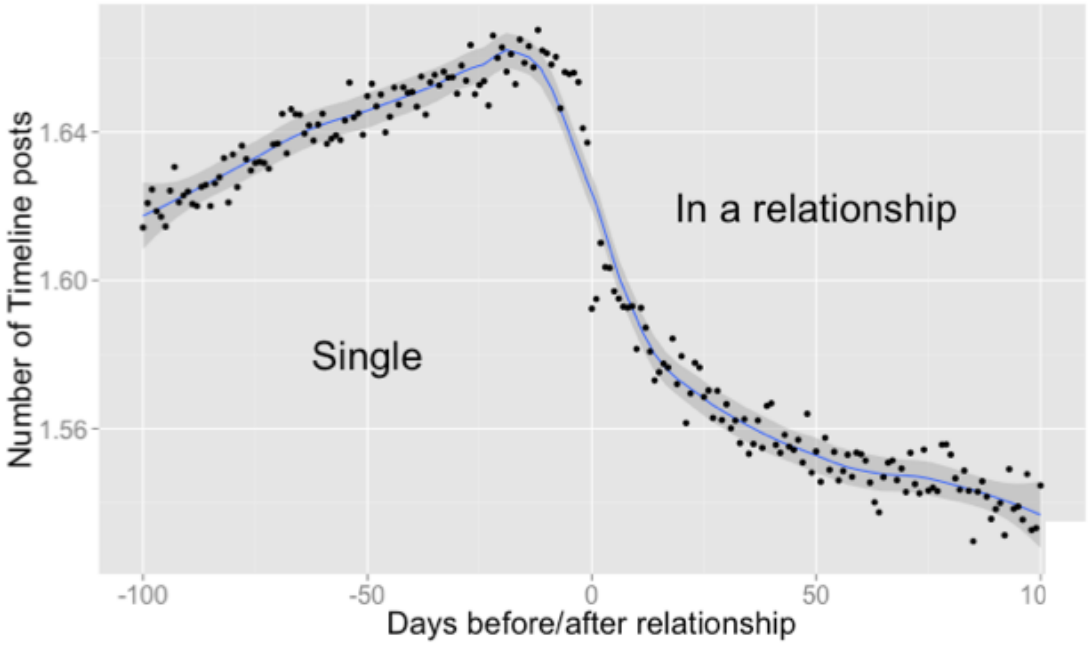
(b) Popularity distribution based on $\Delta rank \cdot \Delta views$



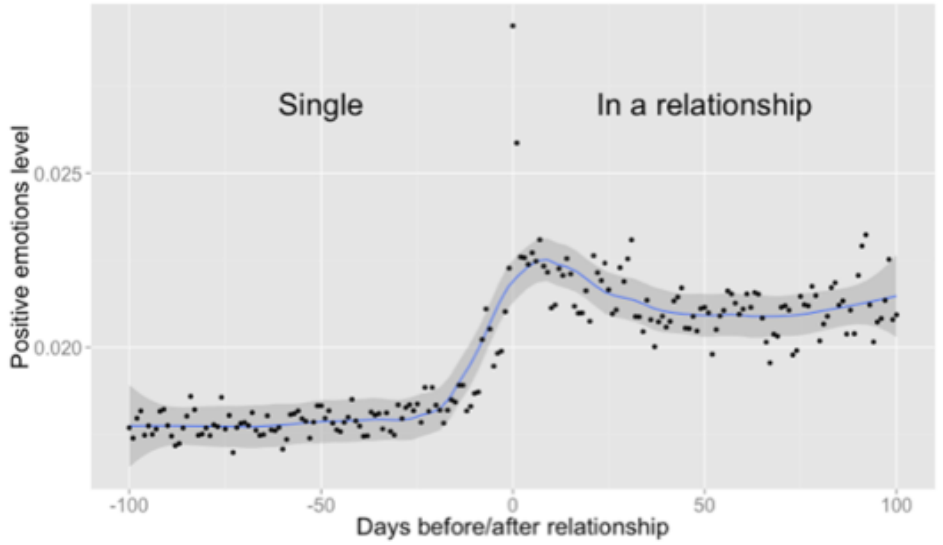
(b) Popularity distribution of videos with varying ages

Figure 9: Changes in ranking and popularity

The Formation of Love



https://www.facebook.com/notes/facebook-data-science/the-formation-of-love/10152064609253859



Growing Closer on Facebook

- Receiving written content—things that take some effort—like comments, posts on a friend's Timeline, and messages are linked to improvements in relationships.

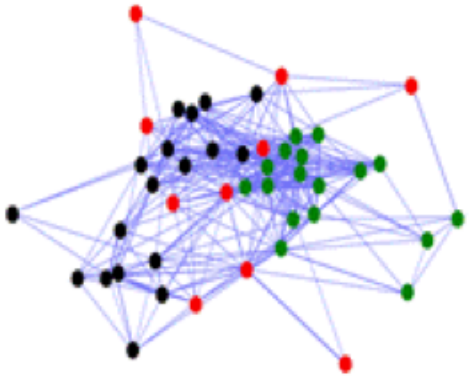
- Likes, on the other hand, have little to no effect on relationships, possibly because Likes "cost less" socially.

- Reading about a friend through feed stories and their Timeline is also associated with growing closer (regardless of whether you talk to that friend online).

- These effects happen *even after you take into account* how often two people talk on the phone, see each other in person, and email. Facebook adds something over these other channels.

- Family relationships aren't affected by Facebook use as much as non-family relationships are. But family members say Facebook helps them see a different side of their relative; Dad becomes a “social person.”


Hamed Haddadi
https://haddadi.github.io



Queen Mary
University of London

Experiment

- Study subjects (n = 46) were taken from a flock of n = 300 merino sheep (Ovis aries) grazed at SARDI, and split into three groups: group A (n=10), group B (n=18) and group C (n=18).
- Sheep were kept in these three groups for 2 weeks in identical sized 0.9-km^2 rectangular fields, and given ad libitum access to hay and water. The three groups were then mixed together.

Location tracking

- All day
- GPS module capable of recording single frequency L1 raw range data at 10 Hz (uBlox LEA-4T GPS module), an IMU comprising a three axis MEMS accelerometer, three axis of MEMS gyroscope and three axis of magnetometer and a GPS patch antenna, MSP430 microcontroller and a rechargeable 2,200 mAh lithium polymer battery.
- Batteries charged overnight

Question: Who's my buddy?

- Using our GPS data, we created spatial matrices detailing the straight line distance between all dyads in the flock
- These data provided us with the total number of seconds sheep i spent at a certain distance from sheep j.
- Preliminary observations of sheep flocking at the site and previous research on Merino sheep suggested that individuals tend to be spaced 1–3 m during normal activity
- calculated adjacency matrices for 30 different spatial–temporal scales that ranged from two individuals spending 1 min at 1 m from one another, to five consecutive minutes at 3.5 m.
- Dyads were therefore defined as 'associating' every time that criterion was met.

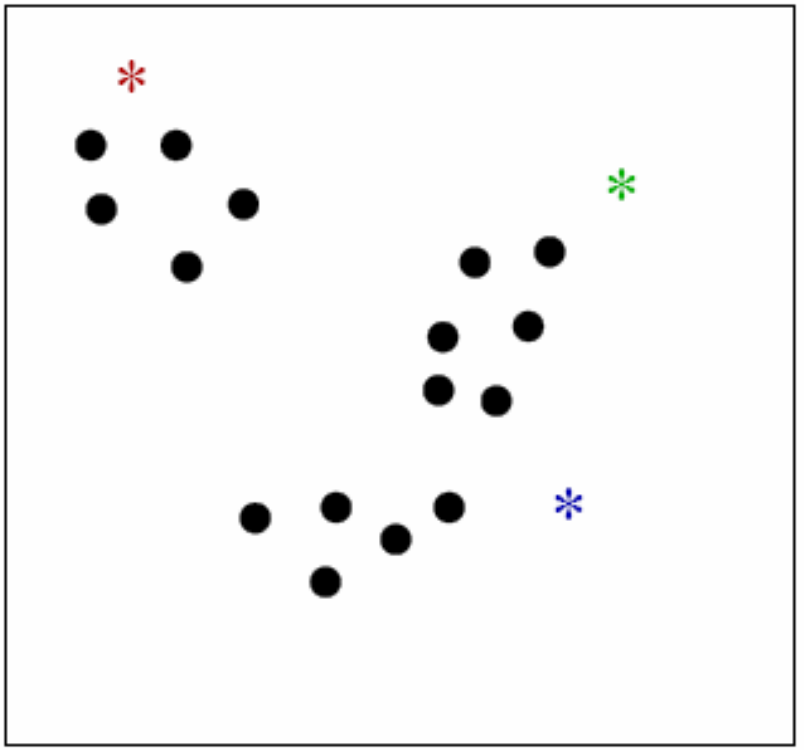
K-Means clustering

- K = # of clusters (given); one "mean" per cluster
- Interval data

- Initialize means (e.g. by picking k samples at random)
- Iterate:

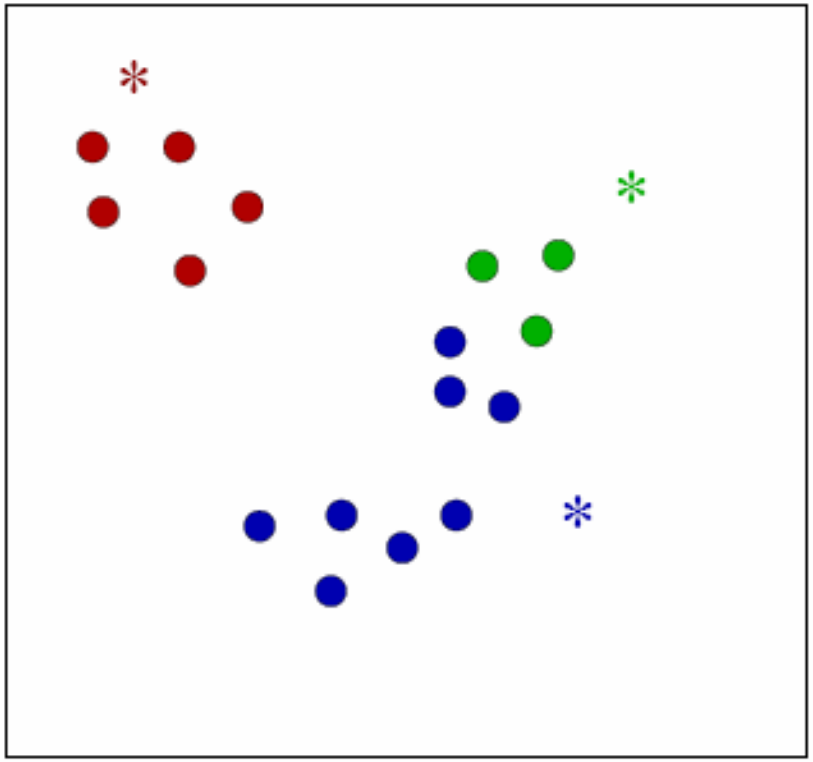
(1) assign each point to nearest mean

(2) move "mean" to center of its cluster.

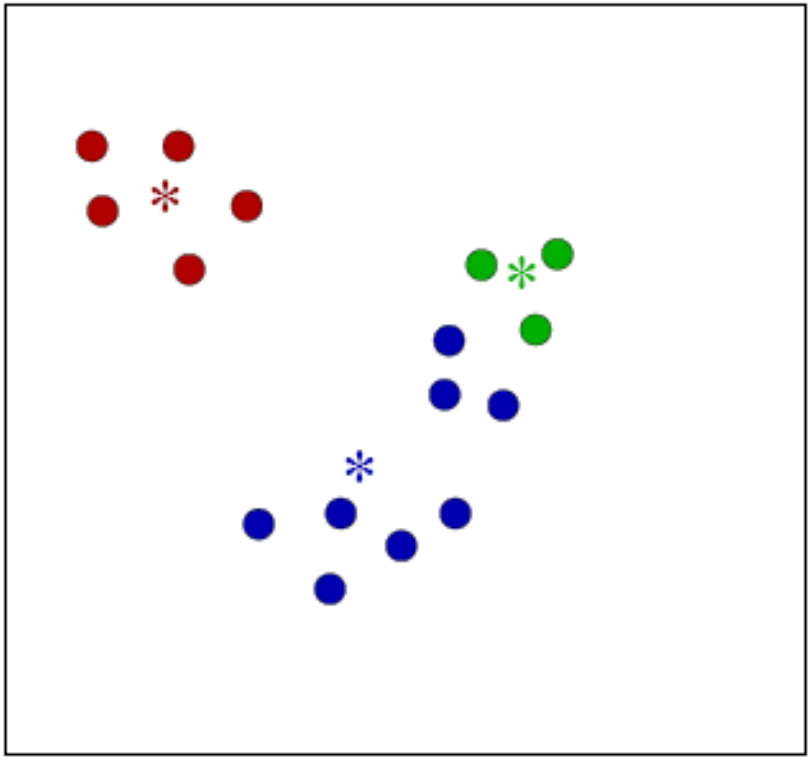


Initialize representatives ("means")

Assignment Step; Means Update



Assign to nearest representative

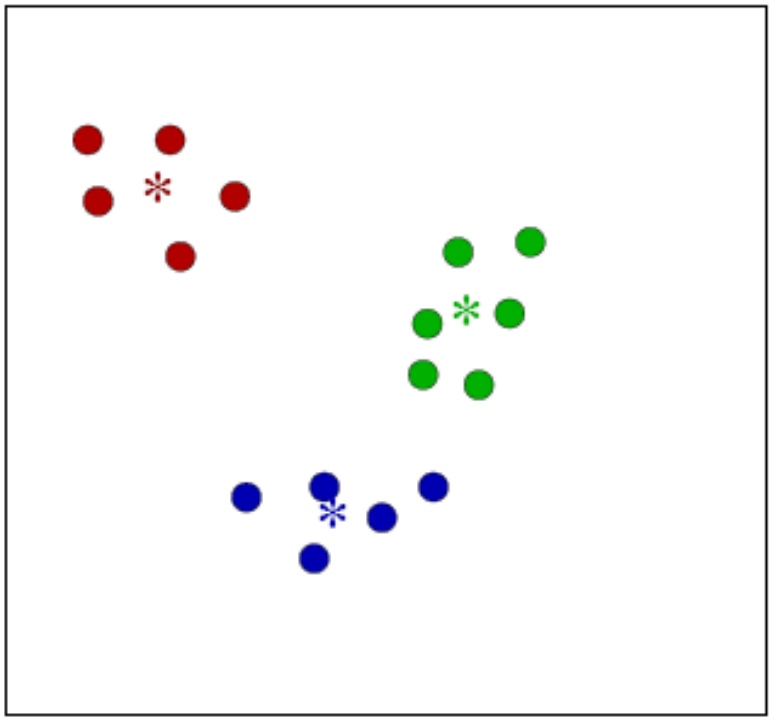


Re-estimate means

Convergence after another iteration

Complexity:

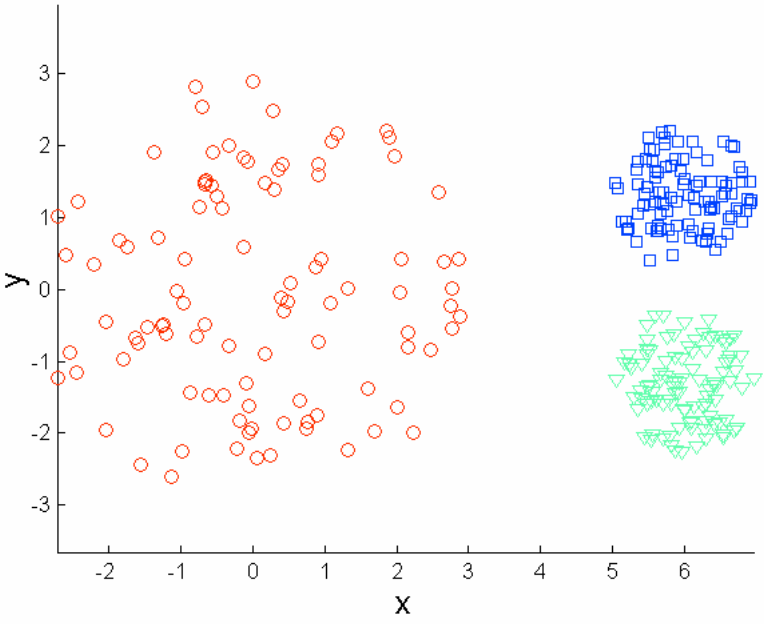
O(k . n . # of iterations



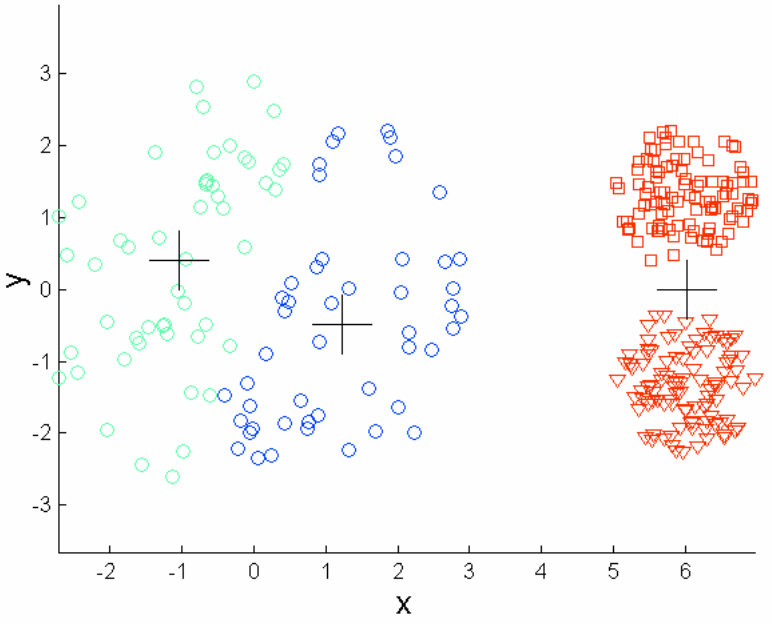
The objective function is

$$\min_{\{\boldsymbol{\mu}_1,\cdots,\boldsymbol{\mu}_k\}} \sum_{h=1} \sum_{\mathbf{x}\in\mathcal{X}_h} \|\mathbf{x}-\boldsymbol{\mu}_h\|^2$$

Limitations of K-means: Differing Density

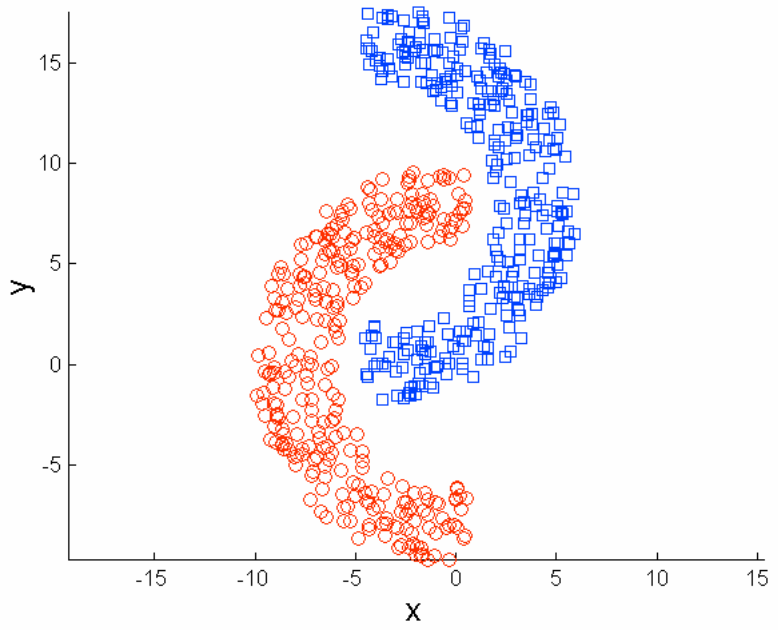


Original Points

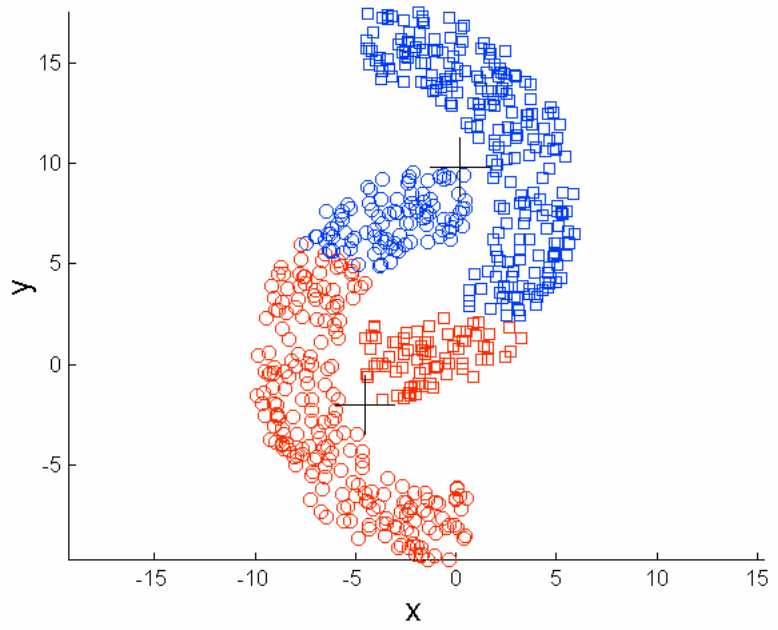


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes

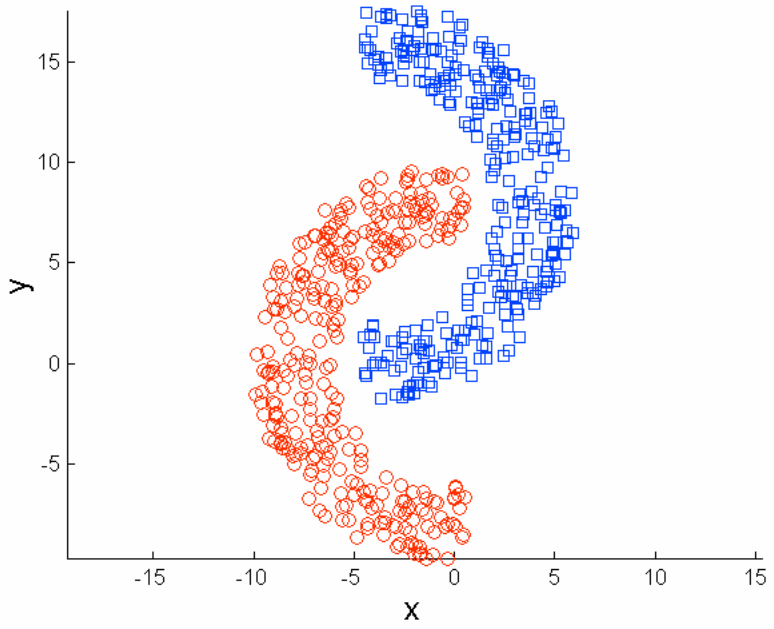


Original Points

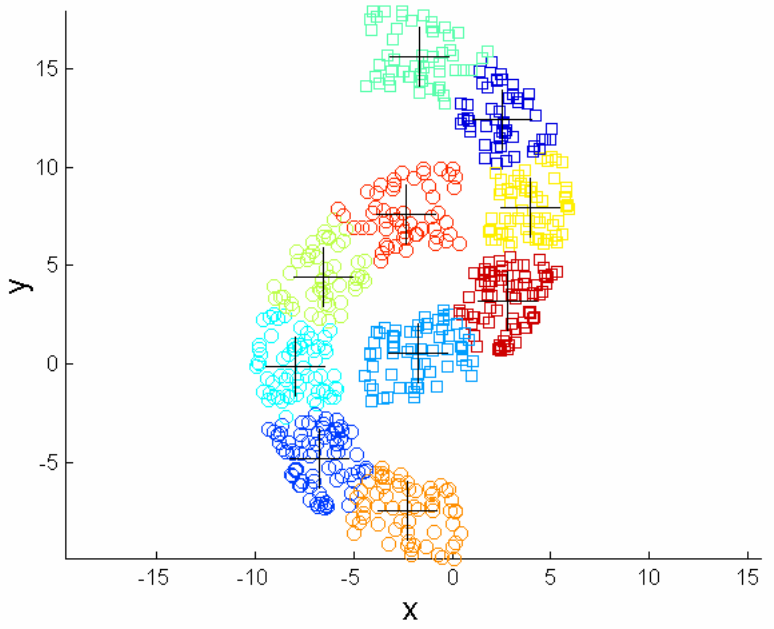


K-means (2 Clusters)

Overcoming K-means Limitations



Original Points



K-means Clusters

Spatial–temporal criterion

Table 1 Performance of k-means in detecting familiar individuals once mixed together into one larger flock at 30 different spatial–temporal scales

		Distance (meters)					
		1	1.5	2	2.5	3	3.5
Time (minutes)	1	0.56	0.58	0.60	0.59	0.85	0.79
	2	0.49	0.57	0.46	0.58	0.64	0.72
	3	0.37	0.55	0.60	0.85	0.69	0.70
	4	0.52	0.58	0.81	0.80	0.70	0.70
	5	0.51	0.56	0.57	0.81	0.80	0.75

Warmer colours in the plot represent higher accuracy

Entropy rate

We calculate the network graph entropy rate—a measure of ease of spreading of information (e.g. a disease) in a network—for every second of our dataset to determine the minimum sampling rate required to capture the variability observed in our sheep networks during distinct activity phases.

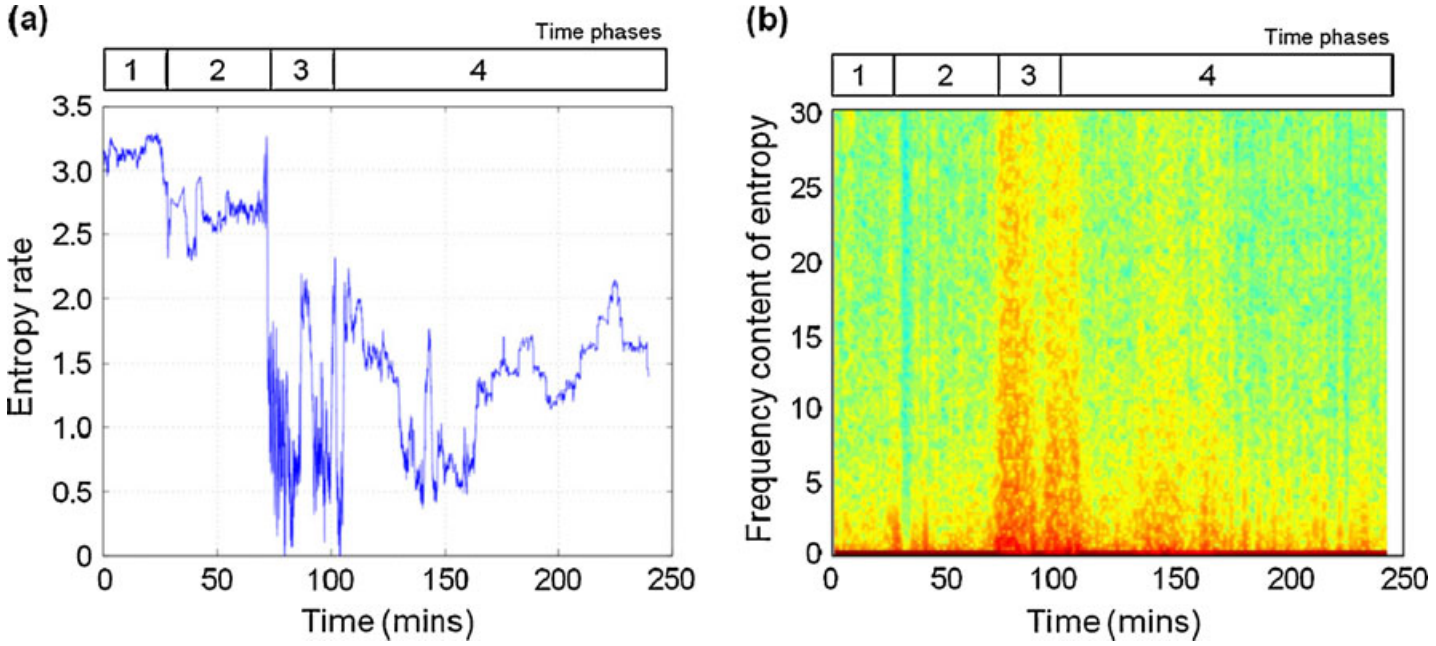
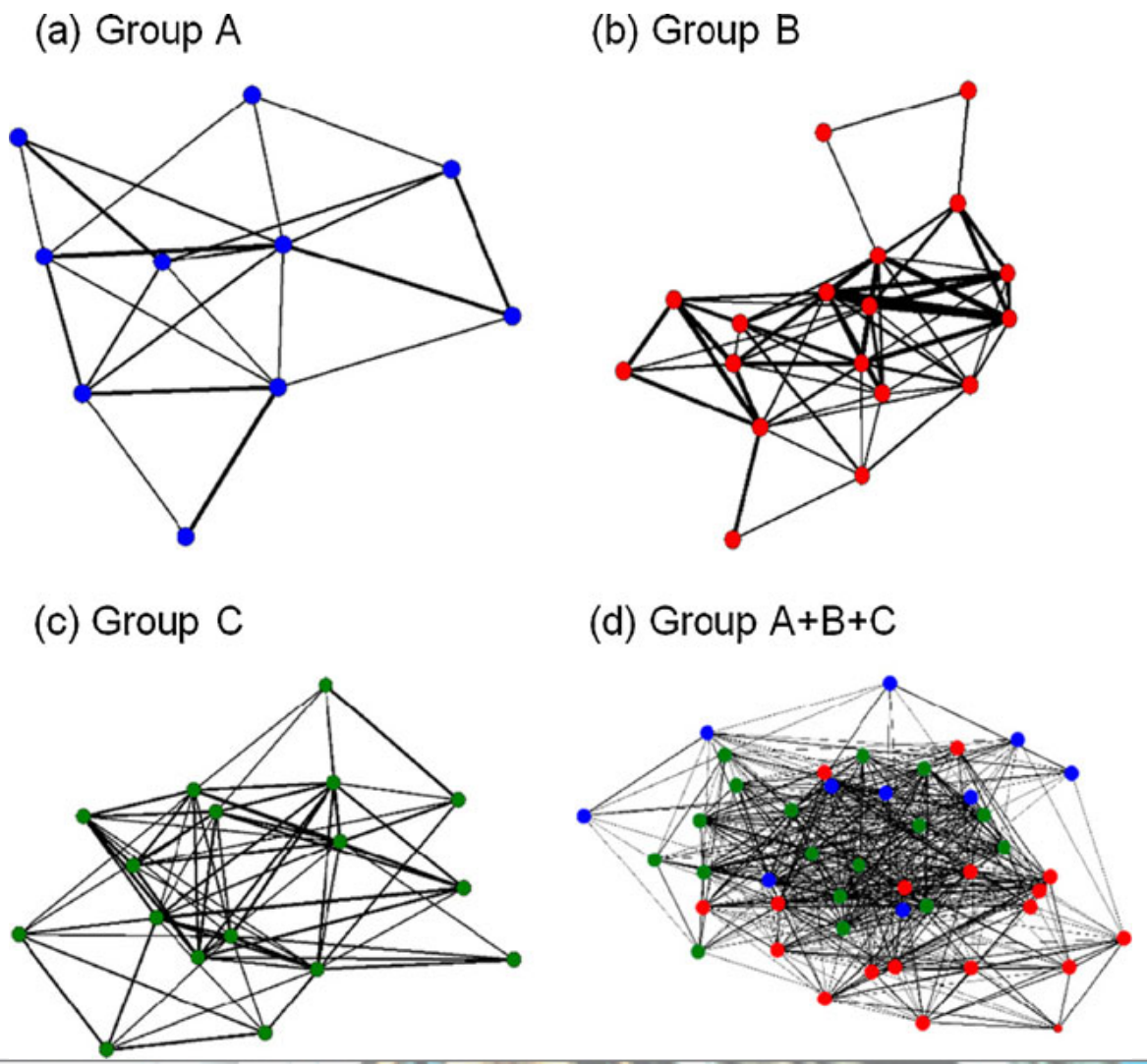


Fig. 4 **a** Entropy rate over time and **b** frequency content of entropy (a spectrogram) depicting the changing structure of the flock over time. In each figure, four distinct activity periods are labelled as time phases: 1='holding pen'; 2='herding'; 3='entry into field'; 4='in field' (see Methods for more details). At low entropy, rates indicate the flock is very dispersed, while a high entropy rate indicates a highly associated flock (also see Fig. 6). The spectogram shows that the frequency content illustrates the highly variable structure of the flock (*warm colours*) during phase 3 when the sheep flock enters the novel new field

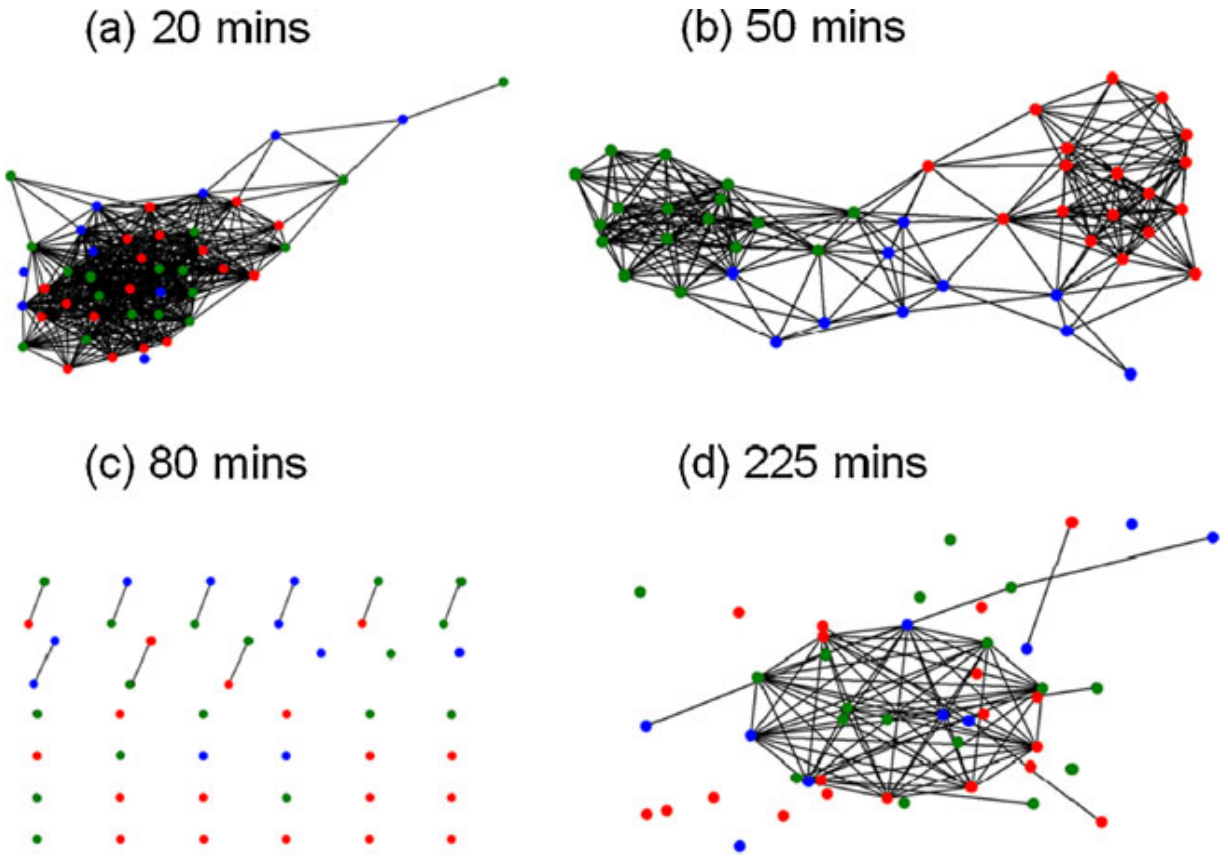
Mixing it up!

Fig. 3 Sociograms depicting spatial associations of the three individual groups before mixing **a**, **b** and **c**, and the mixed group on day of mixing **d**. In **d**, *blue nodes* represent sheep in group A, *green nodes* represent sheep in group C and *red nodes* represent sheep in group B. In all cases, the thickness of the *lines* (edges) indicates the frequency of associations between each dyad, and the network is filtered so that only links above the group mean average are shown for ease of illustration



Bonding

Fig. 5 Sociograms depicting spatial associations of the mixed group taken at four different single second 'snapshots' for a newly formed sheep flock. Nodes represent individual sheep and *lines* (edges) indicate an association between dyads at 2.5 m. Each network's corresponding entropy rate can be seen in Fig. 5a



https://www.youtube.com/watch?v=o54zO30lnas

References

- Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee. 2007. Measurement and analysis of online social networks. In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement (IMC '07)
- Burke, M., and Kraut, R. (to appear). Growing Closer on Facebook: Changes in Tie Strength Through Social Network Site Use. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2014.
- Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. 2007. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In Proceedings of the 7th ACM SIGCOMM conference on Internet measurement
- Hamed Haddadi, Andrew J. King, Alison P. Wills, Damien Fay, John Lowe, A. Jennifer Morton, Stephen Hailes, and Alan Wilson, "Determining association networks in social animals: choosing spatial-temporal criteria and sampling rates", in Behavioral Ecology and Sociobiology
- Meeyoung Cha, Juan Antonio Navarro Perez, Hamed Haddadi, "The Spread of Media Content Through Blogs", in Social Network Analysis and Mining, Springer, Volume 2, Number 3, (September 2012)
- Meeyoung Cha, Fabrício Benevenuto, Hamed Haddadi and Krishna Gummadi, “The world of connections and information flow in Twitter”, IEEE Transactions on Systems, Man and Cybernetics - Part A (Impact Factor: 2.54). Volume 42, Number 4. July 2012.