them, and let them move. Similar records will move closer to one another, pulling their neighbors with them (and thus providing transitive closure). We query the system by inserting a query record that contains what we know about the POI, letting the agents move until the system has converged, and retrieving records that end up close to the query. The closer a record is to the query, the higher we rank it in our list of persons to contact.

We emulate the features of physical movement.

- Extremely close agents repel one another, keeping similar records from collapsing to the same location.
- Interaction decreases with distance, so most interactions are local. Local interactions reduce the set of agents with which a given agent effectively interacts, allowing their influence to be felt in fewer steps.
- The concept of multiple forces lets us handle heterogeneous records with varying field contents. Integration of these forces through agent movement allows transitive interactions among records whose fields do not directly overlap.

Our implementation, discussed in detail in the full paper, includes similarity computation, force definition, distributed execution, and convergence detection. The full paper also discusses the relation between FBC and other technologies, including semantic analysis, cluster analysis, and multi-dimensional scaling.

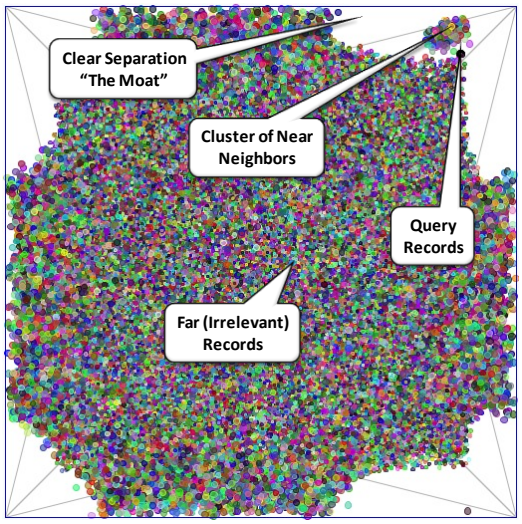

Figure 2: Converged Location of all 350k Records in Clustering Space

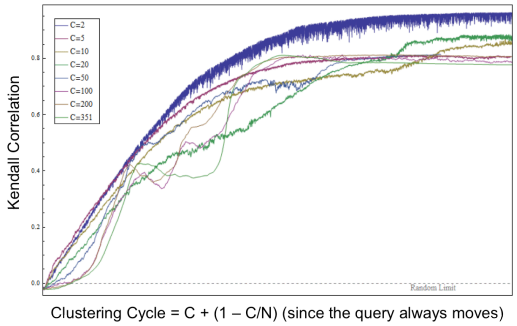

Figure 1: Kendall Correlation over Clustering Cycle * *C*.

EVALUATION.—FBC is inherently parallel and can be distributed for essentially linear (with the number of processors) performance gains over large scale networks and potentially deployed into a MapReduce/Hadoop cloud-computing environment. Our experiment used three standard WinTel PCs to execute 4 clustering processes each and an additional PC to run the MySQL database with the 350k records and their clustering coordinates. In this small setup, we arrived at the results reported here in less than two days execution even though one PC (4 processes) failed due to network problems after less than 8 hours. The clustering space is a unit (1x1x1) box with all 6 faces wrapped. Figure 2 shows the raw result. It highlights the common location of the three query records (phone, address, phone+address) in the upper right corner. Adjacent to the queries, the information matching process highlights a relatively small set of nearby neighbors (data records) as relevant, clearly separated by a "Moat" from the rest of the data. Analysis of this result in the full paper shows that

- The process is highly selective, discarding records that are superficially similar;
- The mass of irrelevant records does contain structure that can support parallel queries;
- The cluster of near neighbors includes records with no direct similarity, supporting transitive closure of the process;
- We found the person of interest, as later confirmed by the sponsor who provided the data.

We assessed the convergence of FBC with an artificial data set of 350 color (RGB) data records, groups into seven clusters, and one query record. We start the experiment with a random arrangement of the records' agents in cluster space and run to (manually determined) convergence, for various values of the parameter *C* that determines how many randomly selected agents are allowed to interact in each cycle. We assess the quality of clustering using Kendall correlation. Figure 1 shows the exponential shape of convergence. It also assesses the impact of parallelization by scaling the x-axis for each data series by a factor of *C* and correcting for the movement of the query record. Thus scaled, the convergence curves trend very close to each other, suggesting a nearly linear speed-up with the number of processors.

CONCLUSIONS.—The current project demonstrates the ability of FBC to find transitively related groups of records in a distributed environment that can scale to handle massive data. The full paper discusses opportunities for extension, including more disciplined *weighting* of different similarity components, provision for further *human interaction*, *distributing data* as well as processing, and running with *dynamically changing data*.

Many important applications in epidemiology and domestic security require the ability to discover transitive linkages across heterogeneous databases rapidly, without reasoning explicitly about possible scenarios. Instead of reasoning about the various records, Force-Based Clustering (FBC) turns each record into a software agent that moves in an abstract information space in response to the net "force" it feels from other agents. These forces in turn are defined by generic similarity measures over commonly occurring fields, measures that can readily be defined in advance and applied quickly to available information. The agent interactions can be distributed over many processors to speed the clustering process. Application of this approach to a synthetic data set (provided by an anonymous sponsor external to our research group) allows us to identify the person of interest.

The full paper is available at
http://abcresearch.org/papers/AAMAS2013TIM.pdf.