Tracking STARS in the Universe

Corinna Cortes and Daryl Pregibon
AT&T Labs - Research

Abstract

We draw certain analogies between computational astronomy and mega-business applications as regards the ability to track large numbers of objects or entities. In astronomy the objects might be stars while in the world of telecommunications, STARS might be Standard Telephone Alphanumeric Recording Symbols (i.e. telephone numbers). In both cases the universe contains many such entities (trillions of stars and billions of STARS).

Two challenging problems concern the size of the data stream (numbering in the hundreds of millions of sightings/transactions per day) and the need to segment (e.g., classify or catalog) the objects according to the type of star/STAR. We discuss visualization issues that allow analysts to inspect the quality of an inferred segmentation on both the macro level (clusters of STARS) and the micro level (individual STARS).

1 Background

Due to regulatory and technological changes in the past 15 years, the telecommunications arena has transformed itself from a monopolistic industry to one of the most competitive businesses in the United States. In such a market place, understanding customer needs and usage patterns is essential, not only for introducing new services, but in assessing the competitive landscape.

We propose a methodology for characterizing customer behavior based on large-scale tracking. The raw material for tracking consists of network data. AT&T carries over 200 million phone calls per day on its domestic telecommunications network. The calls can be voice, data, or fax transmissions. Every call generates a record (for billing purposes) consisting of five basic fields

- the originating telephone number
- the terminating telephone number
- the date
- the time of the call
- the duration of the call

The simplicity of these records is compensated by their number: collecting, moving, analyzing, and storing 200 million such records per day requires serious computing. Exacerbating the challenge is that business needs dictate that we understand each and every STAR, in the very least, for cataloging purposes.

Network data does not provide customer identification, only pairs of phone numbers (denoted in this paper by the term STARS). We propose to track the behavior of STARS by simply scoring them according to the conditions under which they were sighted. The next section illustrates that these conditions contain a very strong signal that can be exploited in a statistical monitoring algorithm. A feature of the proposed methodology is that it allows for changing behavior through time, as would be expected as lifestyle changes occur, as well as the more dramatic change when a STAR is reassigned to a new customer.

2 Methodology for a specific application: Bizosity

Customer care organizations are particularly interested in segmenting all observed STARS according to account type (either as a business or a residence). One leading application is where calls coming into the care channel can be routed (by call-center software) to the appropriate (business or consumer) customer service representative. This is problematic since the records contain no account information other than the STAR symbol (i.e. 10 digit number).

We focus on this segmentation problem by exploiting statistical variation between known business and residential calling, Fig. 1. Specifically, we exploit the historical tendencies:

- businesses make/receive nearly all their calls between Monday - Friday
- businesses make/receive nearly all their calls between 8am - 6pm
- businesses tend to have short duration calls
- businesses tend to call businesses (and 800 numbers) - residences tend to call other residences (and 800 numbers)

Consider the probability that a STAR is a business given the calling pattern observed over a fixed period of time. This probability can be modeled (off-line) using known business and residence STARS together with observed network data over the fixed time period. The model can then be used to score every STAR for its "bizosity" – how much a STAR *behaves* like a business. The behavioral "bizosity" may be very different from the official status of a STAR, but this is all that can be inferred from network-based data.

Because of the huge calling volume on the AT&T network we opt for a *time-driven* tracking approach, where data is aggregated over one day's worth of calls. The calls are binned according to time-of-day, duration, and termination, Fig. 2, and this vector forms the input to the scoring model. Logistic regression seems to perform quite well for our purpose but due to the near-rank deficiency of the binned data, regularization via a "ridge" penalty is necessitated. Fig. 3 shows the regression coefficients of the model. The model has obviously discovered the known historical patterns of business and residential calling.

We track the evolution of a STAR by blending in todays score with the old (yesterdays) value to produce the new estimate of business-like behavior:

$$BIZOSITY(new) = \theta * BIZOSITY(old) + (1 - \theta) * BIZOSITY(today) \tag{1}$$

where $0 < \theta < 1$ controls the degree to which old data is "aged out". The aging parameter θ has to be small enough that the bizosity score of a STAR will track within reasonable time if the STAR changes behavior, e.g. the STAR gets assigned to a new customer, but large enough that the bizosity does not fluctuate too much just because the STAR goes on vacation for a week.

The update, Eqn. 1, is done for all recently observed STARS. In particular note that we also update STARS where no calls are observed since the lack of calls also provides information on bizosity (e.g., no calls made on a weekend suggests that a STAR is a business). Overall, we estimate approximately 250 million updates a day.

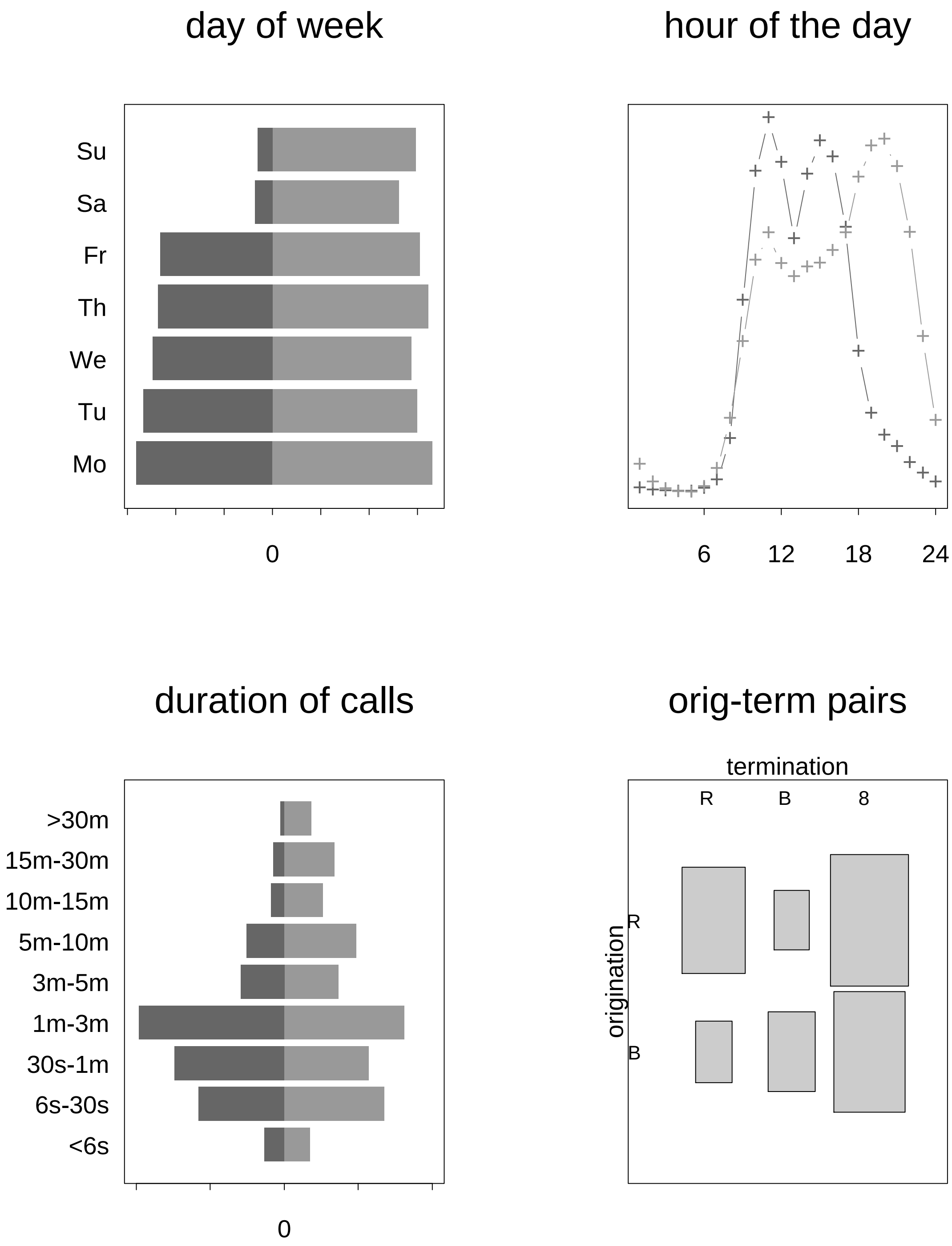

Figure 1: Historical tendencies for business and residential calling. In the first three panels, dark grey represents business calling and light grey represents residential calling. In the fourth panel, the size of the rectangle is proportional to the amount of calls between origination and termination pairs.

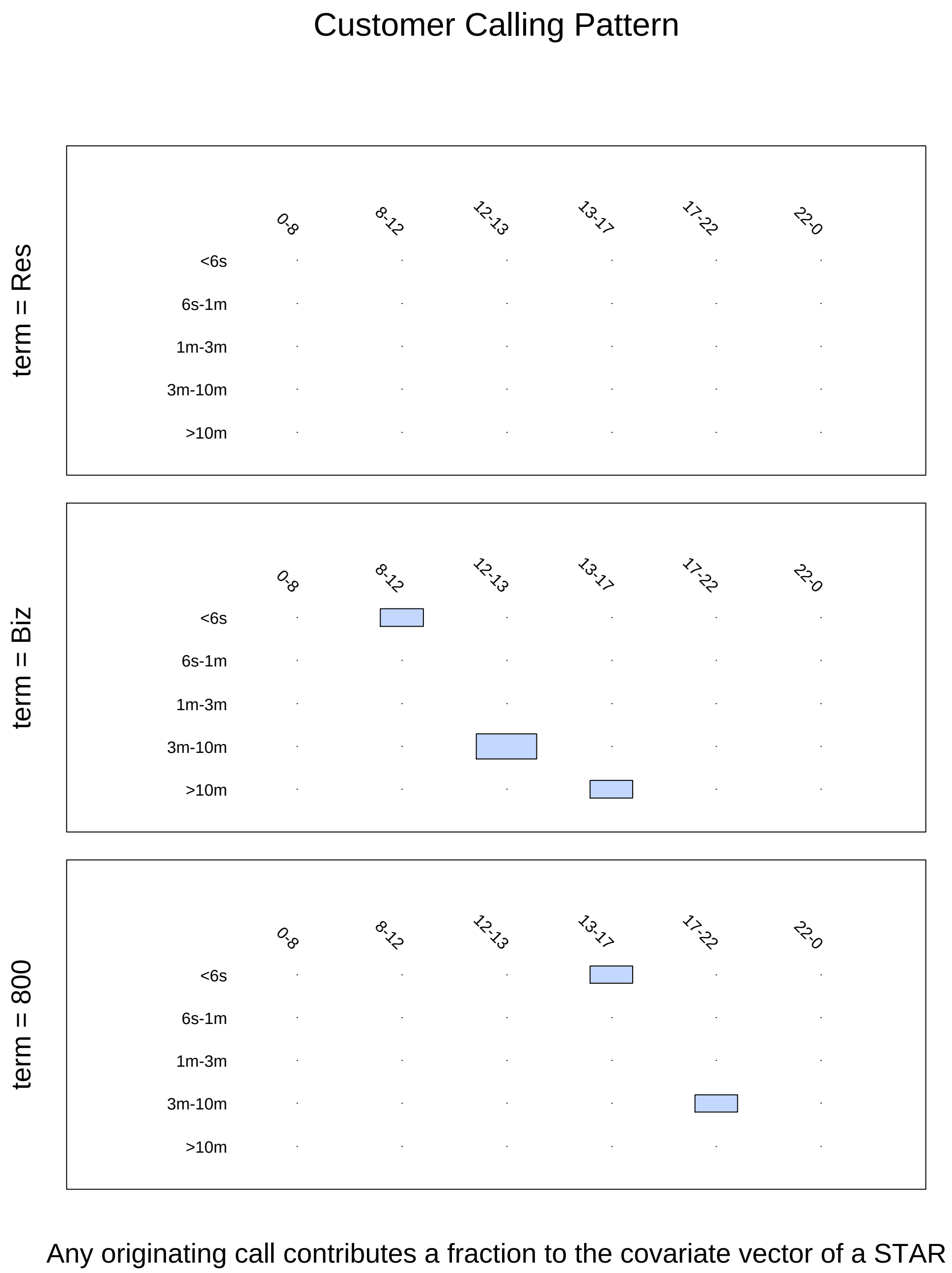

Figure 2: A hypothetical calling pattern for a STAR for one day. This STAR made six calls: four calls into businesses and two calls into 800 numbers at the indicated hours and call durations.

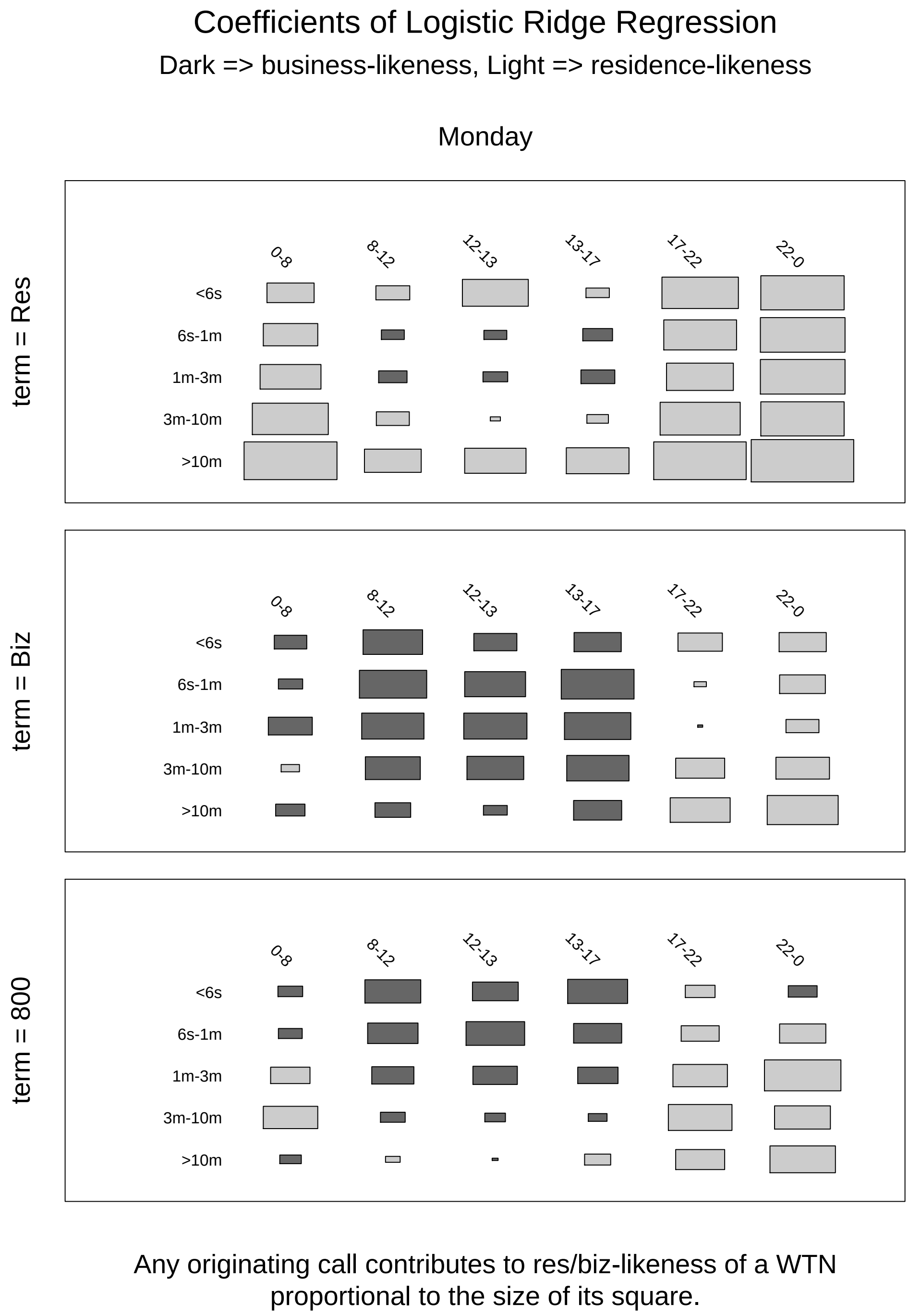

Figure 3: Regression coefficients of the model used to score "bizosity" based on the daily aggregate of calls.

3 System issues

Processing hundreds of millions of calls and bizosity scores every day requires efficient programs and serious hardware. Our processing environment consists of a 20 processor SGI Challenge machine configured with 3 GB of main memory, a similar sized swap space, and close to a terabyte of disk.

Efficiency and robustness are the primary concerns for our software. We have developed a full binary tool box for sorting, filtering, and tabulation so that we minimize (or eliminate completely) expensive ascii conversions. To facilitate robustness we have built data structures that effortlessly deal with growth, for example from area code splits. All data structures are extremely fast for both sequential and random access, but since they exceed 2 GB, it is necessary to run under 64-bit operating systems.

4 Presentation Issues

Implicitly, our methodology involves estimating several hundred million parameters per day. How can one explore such a huge parameter space? Our approach exploits the fact that STARS are aggregated at several levels of a natural hierarchy:

block the group of 100 STARS that share the first 8 digits of the phone number
exchange the group of 10000 STARS that share the first 6 digits of the phone number
area code the group of (<10M) STARS that share the first 3 digits of the phone number
state the group of STARS that reside in the same state (i.e., a group of area codes)
region the group of STARS that reside in the same region of the county (i.e., a group of states)

We make the data available at all these levels through a variety of visualization tools. In effect the tools are designed to be used in a top-down fashion whereby one starts at the highest level of the hierarchy. As more detail is required, the user is able to "drill-down" to obtain it. Depending on the application, information at the STAR level is provided. For example, corporate security applications require direct access to a random-access data warehouse of call records.

For interactive use, we employ a web browser as the user interface. This has proven quite effective in disseminating the results to a broad community of users and platforms over the corporate intranet. All corporate machines reside inside the AT&T firewall and all the sites are password protected to ensure the privacy of STARS.

5 Conclusion

We have prototyped a system for harnessing the vast stream of data for tracking and segmenting billions of STARS. While much remains to be done, we have demonstrated the quality and effectiveness of a dynamic tracking system.