This page is intentionally blank. Proper e-companion title page, with INFORMS branding and exact metadata of the main paper, will be produced by the INFORMS office when the issue is being assembled.

Appendix: the challenge of massive data

From the point of view of traditional statistics, the logistic regression model described in Section 5.1 should always benefit from more data. From a purely theoretical viewpoint, a large sample size N is always a good thing, and theoretical issues arise only when $N < p$. However, in practice, the *estimation* of the model becomes computationally intractable when N is in the millions. We emphasize that the computational challenge arises, not from memory issues (various techniques and software packages, e.g. `biglm` in R, can be used to address that issue), but rather from the estimation of (1), which requires us to optimize an expensive, highly non-linear function.

Recall that (2) is a product of I integrals, where I is the number of unique donor accounts (over 1M in all). Furthermore, each integrand is a product of N_i logistic functions with a normal density, and thus is highly non-linear and non-convex. None of the I integrals has a closed-form solution; consequently, (2) can only be evaluated numerically, e.g. using Monte Carlo integration or Gaussian quadrature. Numerical integration introduces additional error into the evaluation of the likelihood function, and is also expensive for large I and N_i since each integrand must be evaluated multiple times. For these reasons, quadrature methods are infeasible for large problems, leading to both memory and convergence issues for expectation-maximization (EM) algorithms. This issue is well-known in the literature; for example, Karl et al. (2014) finds that EM algorithms scale poorly to large datasets. In our experience, the available computational procedures for solving (3) with random effects simply stalled, crashed, or otherwise failed to produce meaningful results.

With the advent of increasingly large datasets, the statistics literature has now begun to pay closer attention to large-sample data, where N is very large (in the millions) and p is moderately large (several hundred). Even with a large number of samples, such data may be vulnerable to noise accumulation, spurious correlations, and algorithmic instability (Fan et al. 2014). Ideally, statistical methods for such data should be computationally tractable while retaining the theoretical guarantees of classical statistics (such as consistency). In order to scale up to the Red Cross dataset, we synthesize several emerging statistical methodologies, such as small-sample bootstrapping and stability selection, that yield both tractable and rigorous results.

Our approach is based on the idea of "subsampling," or conducting the statistical analysis on a small, randomly generated subset of the large dataset. This is a natural strategy for

dealing with big data, since a small subsample remains, in some sense, representative of the data as a whole. The size M of the subsample can be less than 1% of the total sample size N, allowing us to perform model selection and estimate the mixed-effect model within 1-2 hours.[4] However, if only a single subsample is used, several pitfalls may arise: 1) The statistics literature demonstrates (Bühlmann and Yu 2002) that using a single smaller-order subsample can bias the outcome of model selection by introducing false positives. This can occur if $M < \frac{N}{5}$, which is certainly the case in our application. 2) Furthermore, Bradić (2014) proves the stronger result that, if only a single subsample is considered, it is virtually impossible to retrieve a sparse set of significant features (that is, the probability of doing so vanishes to zero). 3) Subsampling introduces additional noise into the problem. Thus, a single subsample may inflate the variance of the estimated coefficients, analogous to how the variance of a classical sample mean is larger when the sample size is smaller. 4) A feature that appears infrequently in the big data (e.g., the supporter card feature) may be misrepresented in the subsample. Multiple subsamples can give a more representative picture.

To mitigate these issues, we draw S small subsamples, leading to S distinct, independently estimated Lasso models. Each subsample will produce different results: the number of selected features may vary across subsamples, and the set of selected features itself may vary. However, as we describe below, these results can be aggregated to obtain a single final set of accepted features, regression coefficients, and standard errors. Recent work in statistics (Kleiner et al. 2012, 2014, Bradić 2014) proves that, if M and S are correctly chosen, the aggregated results retain theoretical properties such as consistency, and correct bias that may arise with a single subsample.

We separate the estimation procedure into two stages: we perform variable selection first, removing insignificant features to produce a model of reduced size, and then estimate random effects to correct for variation between donors. See e.g. Fan and Li (2012) for a theoretical treatment of an approach separating fixed effect and random effect estimation. Both stages use subsampling to address big data issues.

Model selection. We perform subsampling in line with the technique of Kleiner et al. (2012, 2014) as follows. For each of S subsamples, we draw M communications with replacement from the complete dataset. The work by Kleiner et al. (2014) recommends setting

[4] The cost of estimation is superlinear in the problem size; consequently, if the sample size is reduced by a factor of 100, the computational savings are much greater.

$M = N^\gamma$ for $\frac{1}{2} \leq \gamma < 1$, and obtains robust empirical results for $\gamma = 0.7$. For a dataset with $N = 8.6 \times 10^6$ communications, the size of a single subsample is $M \approx 71,500$. With regard to the number of samples, a common technique (Hastie et al. 2001) is to use $S = \frac{N}{M}$, or approximately $S \approx 120$.

We then perform model selection as in Section 5.1, replacing N by M in (5); however, as long as $M > p$, BIC preserves its theoretical consistency properties, which means that it will still correctly identify significant features (Zhang et al. 2010). To aggregate the results, we use a version of the stability selection criterion of Meinshausen and Bühlmann (2010) as follows. Each subsample $s = 1, ..., S$ produces a different solution λ_s^* of (5), and a different acceptance set $\mathcal{A}(\lambda_s^*)$. Intuitively, the kth feature is more likely to be significant if it is selected by a larger number of these subsets. We include the kth feature in our final model if

$$\frac{1}{S} \sum_{s=1}^{S} 1_{\{k \in \mathcal{A}(\lambda_s^*)\}} \geq \rho, \tag{EC.1}$$

that is, the proportion of samples in which k is selected exceeds a threshold $\frac{1}{2} < \rho < 1$. Note that the extreme cases $\rho = 0$ and $\rho = 1$ correspond to the union and intersection, respectively, of the sets $\mathcal{A}(\lambda_s^*)$. Let $\mathcal{A}^*$ be the set of all k for which (EC.1) holds.

Estimation. To correct for unobserved variation between donors, it is necessary to refit the random effects model of (1) with the additional constraint that $\beta_k = 0$ for $k \notin \mathcal{A}^*$ (as proved in Belloni and Chernozhukov 2013, this also corrects bias in the regression coefficients). However, even with this reduction in the size of the model, (3) remains prohibitively expensive to compute for the entire dataset. Again, we approach this problem through subsampling. To preserve the longitudinal structure of the large dataset across all subsamples, we now use entire panels as the unit of sampling. We modify the BLB technique to include importance sampling from the empirical distribution of the number of communications per panel (shown in Figure 2).

Formally, this is done as follows. Let $M' = I^\gamma$ be the number of donors included in each subsample, and let $S' = \frac{I}{M'}$ be the number of subsamples generated. A single subsample is created by simulating M' realizations of a discrete random variable Z with pmf

$$P(Z = i) = \frac{N_i}{\sum_{i'=1}^{I} N_{i'}}.$$

Let $Z_1, ..., Z_{M'}$ denote these M' sampled values. For each $m' = 1, ..., M'$, if $Z_{m'} = i$, we add N_i communications $y_{i,1}, ..., y_{i,N_i}$ to the subsample. In this way, a particular panel has

a higher probability of being sampled if it contains more communications. Furthermore, if a panel is sampled, we automatically add every communication in that panel to the subsample, thus preserving the longitudinal nature of the data.

It remains to obtain a single set of estimated coefficients from the results of subsampling. We reoptimize (3), subject to $\beta_k = 0$ for $k \notin \mathcal{A}^*$, independently on each of the S' new subsamples. Let $\hat{\beta}_{k,s'}$ be the estimated coefficient of feature k returned by (1) on subsample $s' \in \{1, ..., S'\}$. We calculate

$$\bar{\beta}_k = \frac{1}{S'} \sum_{s'=1}^{S'} \hat{\beta}_{k,s'}$$

and report this as our final estimate of the effect of feature k. In words, we aggregate the results of subsampling by simply averaging the estimated coefficients across subsamples. Under available consistency results for subsampling, this average should converge to the true coefficient β_k with enough subsamples. Then, we let

$$\hat{\sigma}_k^2 = \frac{1}{S'-1} \sum_{s'=1}^{S'} \left(\hat{\beta}_{k,s'} - \bar{\beta}_k \right)^2 \tag{EC.2}$$

be the sample standard error of the regression output across subsamples. We then use $\frac{\bar{\beta}_k}{\hat{\sigma}_k}$ as the relevant t-statistic, with $S'-1$ degrees of freedom, for the null hypothesis that $\beta_k = 0$. Standard techniques can be used to calculate a p-value.

We briefly discuss the choice of (EC.2) to calculate standard errors. Notice that (EC.2) is calculated based only on the estimated coefficients in our subsamples, not on the estimated standard errors produced by the regression model within each subsample. A recent work by Efron (2012) has argued that these within-subsample standard errors do not contribute to the asymptotic standard error of the aggregated estimator $\bar{\beta}_k$. Moreover, Efron (2012) argues that, in fact, (EC.2) over-estimates the true variance, meaning that $\hat{\sigma}_k^2$ will produce conservative confidence intervals. For our purposes, this conservative estimator is sufficient to evaluate the significance of our results.

To summarize, we analyze the massive Red Cross dataset by separating statistical estimation into two stages. The first stage selects the most important features by conducting Lasso-type regularization on each bootstrapped subsample, then aggregating the results with stability selection. The second stage removes all features $k \notin \mathcal{A}^*$, and corrects for the unobserved variation between donors by estimating random effects in this reduced model.

Table EC.1 Deviance residuals of GLMM-Lasso with subsampling vs. plain logistic regression, demonstrated on 10 random subsamples.

Subsample	Plain LR	GLMM-Lasso
1	100.2518	12.30325753
2	41.28834	12.61857265
3	57.07833	12.42619023
4	74.61048	12.8426679
5	86.99473	12.73456152
6	73.44249	12.28891669
7	145.982	12.39356766
8	53.33318	12.2563628
9	79.01672	12.31632096
10	30.35353	12.27578767

In addition to the theoretical advantages of aggregation, we can see from Figures 4(a)-4(b) that our approach empirically produces more conservative feature sets – there is clearly a small core of features that are "agreed" on by a majority of subsamples, but there are also clear outliers in the "tails" of the histograms that are selected in a very small proportion of subsamples (or in just one subsample).

Numerical illustration. We briefly illustrate the advantages of GLMM-Lasso with subsampling over a rougher but simpler technique, namely ordinary logistic regression (LR), in terms of two standard performance metrics (see, e.g., Smithson and Merkle 2014 for details). We compare these methods using 5-fold cross-validation (CV), a common technique in data mining for evaluating the predictive power of a model. First, we compare the deviance residuals achieved by both methods (averaged over the 5 folds in CV). The comparisons are carried out individually on 10 different subsamples, each of size N^{γ}. (As we discussed earlier, it is always necessary to run models on small subsamples in order to tractably obtain results.) The logistic regression model does not perform any model selection; thus, the results illustrate the benefits of using a more parsimonious model with fewer features.

Table EC.1 presents the results of this comparison. Our model outperforms LR (achieves lower deviance) in each subsample. The results are also much more consistent for the aggregated Lasso model (LR fluctuates more across subsamples), suggesting that there is significant benefit in aggregating over multiple subsamples to reduce variance. Recall also from Figures 4(a)-4(b) that aggregation leads to more conservative results: by eliminating outlier features that are not selected by a majority of subsamples, we reduce the risk of over-confidently reporting significance.

Table EC.2 Area under the ROC curve for GLMM-Lasso with subsampling vs. plain logistic regression, demonstrated on 10 random subsamples.

Subsample	Plain LR	GLMM-Lasso
1	0.51984	0.70489
2	0.52615	0.70022
3	0.50530	0.69283
4	0.51175	0.70806
5	0.53812	0.69613
6	0.51591	0.70018
7	0.53985	0.71100
8	0.52010	0.68855
9	0.52511	0.69241
10	0.52399	0.69920

Next, we compare the area under the ROC curve for both methods. This metric is widely used as a measure of accuracy when the data has binary responses with a small proportion of 1s (as is the case in our application). Results for 10 subsamples are given in Table EC.2. The Lasso model consistently outperforms LR (achieves higher AUC). Furthermore, LR generally has poor predictive power (AUC close to 0.5).

These results are quite consistent with what is known about Lasso in the literature. Classical models, such as logistic regression, estimate coefficients for a large number of features that Lasso simply removes from the model. Consequently, any prediction made by such models is subject to a much higher level of noise. Even if a plain LR model were to accurately estimate some of the coefficients, these accurate estimates are essentially drowned out by a large number of inaccurate estimates for other features. This issue, known as "noise accumulation," is quite common; for example, Fan et al. (2014) discusses how the performance of LR is often no better than random guessing in the presence of noisy data. Furthermore, simple linear models may produce over-inflated standard errors when the data is subject to a high degree of empirical correlation, an issue discussed in Section 6.3. In such settings, the p-values produced by LR may themselves be unreliable (Schaefer 1986), while Lasso is known to be less vulnerable to this issue. These examples illustrate the benefits offered by model selection in analyzing large datasets.

References

Belloni, A., V. Chernozhukov. 2013. Least squares after model selection in high-dimensional sparse models. *Bernoulli* **19**(2) 521–547.

Bradić, J. 2014. Support recovery via weighted maximum-contrast subagging. Arxiv preprint arXiv:1306.3494v3. `http://arxiv.org/pdf/1306.3494v3.pdf`

Bühlmann, P., B. Yu. 2002. Analyzing bagging. *The Annals of Statistics* **30**(4) 927–961.

Efron, B. 2012. Estimation and accuracy after model selection. *Journal of the American Statistical Society* **109**(507) 991–1007.

Fan, J., F. Han, H. Liu. 2014. Challenges of big data analysis. *National Science Review* **1**(2) 293–314.

Fan, Y., R. Li. Variable selection in linear mixed effects models. 2012. *The Annals of Statistics* **40**(4) 2043–2068.

Hastie, T., R. Tibshirani, J. Friedman. 2001. *The elements of statistical learning*. Springer, New York, NY.

Karl, A. T., Y. Yang, S. L. Lohr. 2014. Computation of maximum likelihood estimates for multiresponse generalized linear mixed models with non-nested, correlated random effects. *Computational Statistics and Data Analysis* **73** 146–162.

Kleiner, A., A. Talwalkar, P. Sarkar, M. I. Jordan. 2012. The big data bootstrap. *Proceedings of the 29th International Conference on Machine Learning* 1759–1766.

Kleiner, A., A. Talwalkar, P. Sarkar, M. I. Jordan. 2014. A scalable bootstrap for massive data. *Journal of the Royal Statistical Society* **B76**(4) 795–816.

Meinshausen, N., P. Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society* **B72**(4) 417–473.

Schaefer, R. L. 1986. Alternative estimators in logistic regression when the data are collinear. *Journal of Statistical Computation and Simulation* **25**(1-2) 75–91.

Smithson, M., E. C. Merkle. 2014. *Generalized linear models for categorical and continuous limited dependent variables*. Chapman & Hall/CRC.

Zhang, Y., R. Li, C. L. Tsai. 2010. Regularization parameter selections via generalized information criterion. *Journal of the American Statistical Association* **105**(489) 312–323.