Estimating Traffic and Anomaly Maps via Network Tomography†

Morteza Mardani and Georgios B. Giannakis (contact author)∗

***Abstract*—Mapping origin-destination (OD) network traffic is pivotal for network management and proactive security tasks. However, lack of sufficient flow-level measurements as well as potential anomalies pose major challenges towards this goal. Leveraging the spatiotemporal correlation of nominal traffic, and the sparse nature of anomalies, this paper brings forth a novel framework to map out nominal and anomalous traffic, which treats jointly important network monitoring tasks including traffic estimation, anomaly detection, and traffic interpolation. To this end, a convex program is first formulated with nuclear and ℓ_1-norm regularization to effect sparsity and low rank for the nominal and anomalous traffic with only the link counts and a *small* subset of OD-flow counts. Analysis and simulations confirm that the proposed estimator can *exactly* recover sufficiently low-dimensional nominal traffic and sporadic anomalies so long as the routing paths are sufficiently "spread-out" across the network, and an adequate amount of flow counts are randomly sampled. The results offer valuable insights about data acquisition strategies and network scenaria giving rise to accurate traffic estimation. For practical networks where the aforementioned conditions are possibly violated, the inherent spatiotemporal traffic patterns are taken into account by adopting a Bayesian approach along with a bilinear characterization of the nuclear and ℓ_1 norms. The resultant nonconvex program involves quadratic regularizers with correlation matrices, learned systematically from (cyclo)stationary historical data. Alternating-minimization based algorithms with provable convergence are also developed to procure the estimates. Insightful tests with synthetic and real Internet data corroborate the effectiveness of the novel schemes.**

***Index Terms*—Sparsity, low rank, convex optimization, nominal and anomalous traffic, spatiotemporal correlation.**

I. Introduction

Emergence of multimedia services and Internet-friendly portable devices is multiplying network traffic volume day by day [1]. Moreover, the advent of diverse networks of intelligent devices including those deployed to monitor the smart power grid, transportation networks, medical information networks, and cognitive radio networks, will transform the communication infrastructure to an even more complex and heterogeneous one. Thus, ensuring compliance to service-level agreements necessitates ground-breaking management and monitoring tools providing operators with informative depictions of the network state. One such atlas (set of maps) can offer a flow-time depiction of the network origin-destination (OD) flow traffic. Situational awareness provided by such maps will be the key enabler for effective routing and congestion control, network health management, risk analysis, security assurance, and proactive network failure prevention. Acquiring such diagnosis/prognosis maps for large networks however is an arduous task. This is mainly because the number of OD pairs grows promptly as the network size grows, while probing exhaustively all OD pairs becomes impractical even for moderate-size networks [2]. In addition, OD flows potentially undergo anomalies arising due to e.g., cyberattacks and network failures [3], and the acquired measurements typically encounter misses, outliers, and errors.

Towards creating traffic maps, one typically has access to: (D1) link counts comprising the superposition of OD flows per link; these counts can be readily obtained using the single network management protocol (SNMP) [3]; and (D2) *partial* OD-flow counts recorded using e.g., the NetFlow protocol [3]. Extensive studies of backbone Internet Protocol (IP) networks reveals that the nominal OD-flow traffic is spatiotemporally correlated mainly due to common temporal patterns across OD flows, and exhibits periodic trends (e.g., daily or weekly) across time [3]. This renders the nominal traffic having a small intrinsic dimensionality. Moreover, traffic volume anomalies rarely occur across flows and time [3]–[5]. Given the observations (D1) and/or (D2), ample research has been carried out over the years to tackle the ill-posed traffic inference task relying on various techniques that leverage the traffic features as prior knowledge; see e.g., [6]–[13] and references therein.

To date, the main body of work on traffic inference relies on least-squares (LS) and Gaussian [7], [8] or Poisson models [9], and entropy regularization [10]. None of these methods however takes spatiotemporal dependencies of the traffic into account. To enhance estimation accuracy by exploiting the spatiotemporal dependencies of traffic, attempts have been made in [11] and [12]. Using the prior spatial and temporal structures of traffic, [11] applies rank regularization along with matrix factorization to discover the global low-rank traffic matrix from the link and/or flow counts. The model in [11] is however devoid of anomalies, which can severely deteriorate traffic estimation quality. In the context of anomaly detection, our companion work [12] capitalizes on the low-rank of traffic and sparsity of anomalies to unveil the traffic volume anomalies from the link loads (D1). Without OD-flow counts however, the nominal flow-level traffic cannot be identified using the approach of [12].

The present work addresses these limitations by introducing a novel framework that efficiently and scalably constructs network traffic maps. Leveraging recent advances in compressive sensing and rank minimization, first, a novel estimator is put forth, to effect sparsity and low rank attributes for

† Work in this paper was supported by the MURI Grant No. AFOSR FA9550-10-1-0567. Parts of the paper were presented in the *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 26-31, 2013, and in *IEEE Global Signal and Information Processing Workshop*, Austin, Texas, December 3-5, 2013.

∗ The authors are with the Dept. of ECE and the Digital Technology Center, University of Minnesota, 200 Union Street SE, Minneapolis, MN 55455. Tel/fax: (612)626-7781/625-4583; Emails: {morteza,georgios}@umn.edu

the anomalous and nominal traffic components through ℓ_1- and nuclear-norm, respectively. The recovery performance of the sought estimator is then analyzed in the noise-free setting following a deterministic approach along the lines of [14]. Sufficient incoherence conditions are derived based on the angle between certain subspaces to ensure the retrieved traffic and anomaly matrices coincide with the true ones. The recovery conditions yield valuable insights about the network structures and data acquisition strategies giving rise to accurate traffic estimation. Intuitively, one can expect accurate traffic estimation if: (a) NetFlow measures sufficiently many randomly selected OD flows; (b) the OD paths are sufficiently "spread-out" so as the routes form a column-incoherent routing matrix; (c) the nominal traffic is sufficiently low dimensional; and, (d) anomalies are sporadic enough.

Albeit insightful, the accurate-recovery conditions in practical networks may not hold. For instance, it may happen that a specific flow undergoes a bursty anomaly lasting for a long time [4], or certain OD flows may be inaccessible for the entire time horizon of interest with no NetFlow samples at hand. With the network practical challenges however come opportunities to exploit certain structures, and thus cope with the aforementioned challenges. This work bridges this "theory-practice" gap by incorporating the spatiotemporal patterns of the nominal and anomalous traffic, both of which can be learned from historical data. Adopting a Bayesian approach, a novel estimator is introduced for the traffic following a bilinear characterization of the nuclear- and ℓ_1-norms. The resultant nonconvex problem entails quadratic regularizers loaded with inverse correlation matrices to effect structured sparsity and low rank for anomalous and nominal traffic matrices, respectively. A systematic approach for learning traffic correlations from historical data is also devised taking advantage of the (cyclo)stationary nature of traffic. Alternating majorization-minimization algorithms are also developed to obtain iterative estimates, which are provably convergent.

Simulated tests with synthetic network and real Internet-data corroborate the effectiveness of the novel schemes, especially in reducing the number of acquired NetFlow samples needed to attain a prescribed estimation accuracy. In addition, the proposed optimization-based approach opens the door for efficient in-network and online processing along the lines of our companion works in [15] and [16]. The novel ideas can also be applicable to various other inference tasks dealing with recovery of structured low-rank and sparse matrices.

The rest of this paper starts with preliminaries and problem statement in Section II. The novel estimator to map out the nominal and anomalous traffic is discussed in Section III, and pertinent reconstruction claims are established in Section IV. Sections V and VI deal with incorporating the spatiotemporal patterns of traffic to improve estimation quality. Certain practical issues are addressed in Section VIII. Simulated tests are reported in Section IX, and finally Section X draws the conclusions.

Notation: Bold uppercase (lowercase) letters will denote matrices (column vectors), and calligraphic letters will be used for sets. Operators $(\cdot)'$, $\mathrm{tr}(\cdot)$, $\sigma_{\max}(\cdot)$, $[\cdot]_+$, $\oplus$, $\odot$ and $\mathbb{E}[\cdot]$, $\dim(\cdot)$ will denote transposition, matrix trace, maximum singular-value, projection onto the nonnegative orthant, direct sum, Hadamard product, statistical expectation, and dimension of a subspace, respectively; $|\cdot|$ will stand for cardinality of a set, and the magnitude of a scalar. The ℓ_p-norm of $\mathbf{x}\in\mathbb{R}^n$ is $\|\mathbf{x}\|_p := (\sum_{i=1}^n |x_i|^p)^{1/p}$ for $p\geq 1$. For two matrices $\mathbf{M},\mathbf{U}\in\mathbb{R}^{n\times n}$, $\langle \mathbf{M},\mathbf{U}\rangle := \mathrm{tr}(\mathbf{M}'\mathbf{U})$ denotes their trace inner product. The Frobenius norm of matrix $\mathbf{M}=[m_{i,j}]\in\mathbb{R}^{n\times p}$ is $\|\mathbf{M}\|_F := \sqrt{\mathrm{tr}(\mathbf{M}\mathbf{M}')}$, $\|\mathbf{M}\| := \max_{\|\mathbf{x}\|_2=1}\|\mathbf{M}\mathbf{x}\|_2$ is the spectral norm, and $\|\mathbf{A}\|_\infty := \max_{i,j}|a_{ij}|$ the ℓ_∞-norm. The $n\times n$ identity matrix will be represented by $\mathbf{I}_n$ and its i-th column by $\mathbf{e}_i$, while $\mathbf{0}_n$ will stand for the $n\times 1$ vector of all zeros, $\mathbf{0}_{n\times p} := \mathbf{0}_n\mathbf{0}_p'$. Operator vec stacks columns of a matrix, and conversely does unvec; $\cap$ and $\cup$ stand for the set intersection and union, respectively; $\mathrm{supp}(\mathbf{A}) := \{(i,j): a_{ij}\neq 0\}$ is the support set of $\mathbf{A}$, and $[n] := \{1,\ldots,n\}$.

II. Preliminaries and Problem Statement

Consider a backbone IP network described by the directed graph $G(\mathscr{N},\mathscr{L})$, where $\mathscr{L}$ and $\mathscr{N}$ denote the set of links and nodes (routers) of cardinality $|\mathscr{L}|=L$ and $|\mathscr{N}|=N$, respectively. A set of end-to-end flows $\mathscr{F}$ with $|\mathscr{F}|=F$ traverse different OD pairs. In backbone networks, the number of OD flows far exceeds the number of physical links ($F\gg L$). Per OD-flow, multipath routing is considered where each flow traverses multiple possibly overlapping paths to reach its intended destination. Letting $x_{f,t}$ denote the unknown traffic level of flow $f\in\mathscr{F}$ at time t, link $\ell\in\mathscr{L}$ carries the fraction $r_{\ell,f}\in[0,1]$ of this flow; clearly, $r_{\ell,f}=0$ if flow f is not routed through link ℓ. The total traffic carried by link ℓ is then the weighted superposition of flows routed through link ℓ, that is, $\sum_{f\in\mathscr{F}} r_{\ell,f}x_{f,t}$. The weights $\{r_{\ell,f}\}$ form the routing matrix $\mathbf{R}\in[0,1]^{L\times F}$, which is assumed fixed and given. These weights are not arbitrary but must respect the flow conservation law $\sum_{\ell\in\mathscr{L}_{\mathrm{in}}(n)} r_{\ell,f} = \sum_{\ell\in\mathscr{L}_{\mathrm{out}}(n)} r_{\ell,f},\ \forall f\in\mathscr{F}$, where $\mathscr{L}_{\mathrm{in}}(n)$ and $\mathscr{L}_{\mathrm{out}}(n)$ denote the sets of incoming and outgoing links to node $n\in\mathscr{N}$, respectively.

It is not uncommon for some of flow rates to experience sudden changes, which are termed *traffic volume anomalies* that are typically due to the network failures, or cyberattacks [3]. With $a_{f,t}$ denoting the unknown traffic volume anomaly of flow f at time t, the traffic carried by link ℓ at time t is

$$y_{\ell,t} = \sum_{f\in\mathscr{F}} r_{\ell,f}(x_{f,t}+a_{f,t}) + v_{\ell,t},\quad t\in\mathscr{T} \tag{1}$$

where the time horizon $\mathscr{T}$ comprises T slots, and $v_{\ell,t}$ accounts for the measurement errors. In IP networks, link loads can be readily measured via SNMP supported by most routers [3]. Introducing the matrices $\mathbf{Y} := [y_{\ell,t}], \mathbf{V} := [v_{\ell,t}]\in\mathbb{R}^{L\times T}$, $\mathbf{X} := [x_{f,t}]$, and $\mathbf{A} := [a_{f,t}]\in\mathbb{R}^{F\times T}$, link counts in (1) can be expressed in a compact matrix form as

$$\mathbf{Y} = \mathbf{R}(\mathbf{X}+\mathbf{A}) + \mathbf{V}. \tag{2}$$

Here, matrices $\mathbf{X}$ and $\mathbf{A}$ contain, respectively, the *nominal* and *anomalous* traffic flows over the time horizon $\mathscr{T}$. Inferring $(\mathbf{X},\mathbf{A})$ from the compressed measurements $\mathbf{Y}$ is a severely underdetermined task (recall that $L\ll F$), necessitating

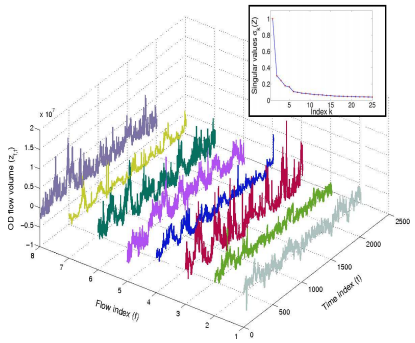


Fig. 1. Internet-2 traffic for a few representative OD flows across time and flows [17].

additional data to ensure identifiability and improve estimation accuracy. A useful such source is the direct flow-level measurements

$$z_{f,t} = x_{f,t} + a_{f,t} + w_{f,t},\ t \in \mathscr{T},\ f \in \mathscr{F} \tag{3}$$

where $w_{f,t}$ accounts for measurement errors. The flow traffic in (3) is sampled via NetFlow [3] at each origin node. This however incurs high cost which means that one can have measurements (3) only for few (f,t) pairs [3]. To account for missing flow-level data, collect the available pairs (f,t) in the set $\Pi \in [F] \times [T]$; introduce also the matrices $\mathbf{Z}_\Pi := [z_{f,t}], \mathbf{W}_\Pi := [w_{f,t}] \in \mathbb{R}^{F\times T}$, where $z_{f,t} = w_{f,t} = 0$ for $(f,t) \notin \Pi$, and associate the sampling operator $\mathscr{P}_\Pi$ with the set Π, which assigns entries of its matrix argument not in Π equal to zero, and keeps the rest unchanged. As with $\mathbf{X}$, it holds that $\mathscr{P}_\Pi(\mathbf{X}) \in \mathbb{R}^{F\times T}$. The flow counts in (3) can then be compactly written as

$$\mathbf{Z}_\Pi = \mathscr{P}_\Pi\left(\mathbf{X} + \mathbf{A}\right) + \mathbf{W}_\Pi. \tag{4}$$

Besides periodicity, temporal patterns common to traffic flows render rows (correspondingly columns) of $\mathbf{X}$ correlated, and thus $\mathbf{X}$ exhibits a few dominant singular values which make it (approximately) low rank [3]. Anomalies on the other hand are expected to occur occasionally, as only a small fraction of flows are supposed to be anomalous at any given time instant, which means $\mathbf{A}$ is sparse. Anomalies may exhibit certain patterns e.g., failure at a part of the network may simultaneously render a subset of flows anomalous; or certain flows may be subject to bursty malicious attacks over time.

Given the link counts $\mathbf{Y}$ obeying (2) along with the partial flow-counts $\mathbf{Z}_\Pi$ adhering to (4), and with $\{\mathbf{R}, \Pi\}$ known, this paper aims at accurately estimating the unknown *low-rank* nominal and *sparse* anomalous traffic pair $(\mathbf{X}, \mathbf{A})$.

III. Maps of Nominal and Anomalous Traffic

In order to estimate the unknowns of interest, a natural estimator accounting for the low rank of $\mathbf{X}$ and the sparsity of $\mathbf{A}$ will be sought to minimize the rank of $\mathbf{X}$, and the number of nonzero entries of $\mathbf{A}$ measured by its ℓ_0-(pseudo) norm. Unfortunately, both rank and ℓ_0-norm minimization problems are in general NP-hard [18]–[20]. The nuclear-norm $\|\mathbf{X}\|_* := \sum_k \sigma_k(\mathbf{X})$, where $\sigma_k(\mathbf{X})$ signifies the k-th singular value of $\mathbf{X}$, and the ℓ_1-norm $\|\mathbf{A}\|_1 := \sum_{f,t} |a_{f,t}|$ are typically adopted as *convex* surrogates [19], [20]. Accordingly, one solves

$$\text{(P1)}\quad (\hat{\mathbf{X}}, \hat{\mathbf{A}}) = \arg\min_{(\mathbf{X},\mathbf{A})} \frac{1}{2}\|\mathbf{Y} - \mathbf{R}(\mathbf{X}+\mathbf{A})\|_F^2 + \frac{1}{2}\|\mathscr{P}_\Pi(\mathbf{Z} - \mathbf{X} - \mathbf{A})\|_F^2 + \lambda_*\|\mathbf{X}\|_* + \lambda_1\|\mathbf{A}\|_1$$

where $\lambda_1, \lambda_* \geq 0$ are the sparsity- and rank-controlling parameters. From a network operation perspective, the estimate $\hat{\mathbf{A}}$ maps out the network health-state across both time and flows. A large value $|\hat{a}_{f,t}|$ indicates that at time instant t flow f exhibits a sever anomaly, and therefore appropriate traffic engineering and security tasks need to be run to mitigate the consequences. The estimated map of nominal traffic $\hat{\mathbf{X}}$ is also a viable input for network planning tasks.

From the recovery standpoint, (P1) subsumes several important special cases, which deal with recovery of $\hat{\mathbf{X}}$ and/or $\hat{\mathbf{A}}$. In the absence of flow counts, i.e., $\Pi = \emptyset$, exact recovery of the *sparse* anomaly matrix $\hat{\mathbf{A}}$ from link loads is established in [12]. The key to this is the sparsity present, which enables recovery from compressed linear-measurements. However, the (possibly huge) nullspace of $\mathbf{R}$ challenges identifiability of the nominal traffic matrix $\mathbf{X}$, as will be delineated later. Moreover, with only flow counts partially available, (P1) boils down to the so-termed robust principal component pursuit (PCP), for which exact reconstruction of the *low-rank* nominal traffic component is established in [14]. Instrumental role in this case is played by the dependencies among entries of the low-rank component, reflected in the observations. Indeed, the matrix of anomalies is not recoverable since observed entries do not convey any information about the unobserved anomalies. Furthermore, without the sparse matrix, i.e., $\mathbf{A} = \mathbf{0}$, and only with flow counts partially available, (P1) boils down to the celebrated matrix completion problem studied e.g., in [21], which can be applied to interpolate the traffic of unreachable OD flows from the observed ones at the edge routers.

The aforementioned considerations regarding recovery in these special cases make one hopeful to retrieve $\mathbf{X}$ and $\mathbf{A}$ via (P1). Before delving into the analysis of (P1), it is worth noting that [22] has recently studied recovery of compressed low-rank-plus-sparse matrices, also known as compressive PCP, where the compression is performed by an orthogonal projection onto a low-dimensional subspace, and the the support of the sparse matrix is presumed uniformly random. The results require certain subspace incoherence conditions to hold, which in the considered traffic estimation task impose strong restrictions on the routing matrix $\mathbf{R}$ and the sampling operator $\mathscr{P}_\Pi(\cdot)$. Furthermore, it is unclear how to relate the subspace incoherence conditions to the well-established incoherence measures adopted in the context of matrix completion and compressive sampling, which are satisfied by various classes of random matrices; see e.g., [20], [23].

Before closing this section, it is important to recognize that albeit few the NetFlow measurement $\mathbf{Z}_\Pi$, they play an important role in estimating $\mathbf{X}$. In principle, if one merely knows the link counts $\mathbf{Y}$, it is impossible to accurately identify $\mathbf{X}$ when the only prior information about $\mathbf{X}$ and $\mathbf{A}$ is that they are sufficiently low-rank and sparse, respectively. This

identifiability issue is formalized in the next lemma.

Lemma 1: *With $\mathscr{N}_R$ denoting the nullspace of $\mathbf{R}$, and $\mathbf{X}_0 = \mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0^{'}$, if $\mathscr{N}_R \neq \emptyset$, and one only knows $\{\mathbf{Y}, \mathbf{R}\}$, then for any $\mathbf{W} \in \mathscr{N}_R$ the matrix pair $\{\mathbf{X}_1 := \mathbf{X}_0 + \mathbf{W}\mathbf{V}_0', \mathbf{A}_0\}$: (i) is feasible, and (ii) it satisfies $\mathrm{rank}(\mathbf{X}_1) \leq \mathrm{rank}(\mathbf{X}_0) =: r$.*

Proof: Clearly (i) holds true since $\mathbf{RW} = \mathbf{0}$, and subsequently $\mathbf{R}(\mathbf{A}_0 + \mathbf{X}_1) = \mathbf{R}(\mathbf{A}_0 + \mathbf{X}_0) + \mathbf{RW}\mathbf{V}_0' = \mathbf{Y}$. Also, (ii) readily follows from Sylvester's inequality [24] which implies that $\mathrm{rank}(\mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0^{'} + \mathbf{W}\mathbf{V}_0^{'}) \leq \min\{\mathrm{rank}(\mathbf{X}_0 + \mathbf{W}\mathbf{V}_0^{'}), \mathrm{rank}(\mathbf{V}_0)\} \leq \mathrm{rank}(\mathbf{V}_0) = r$. ■

IV. Reconstruction Guarantees

This section studies the exact reconstruction performance of (P1) in the absence of noise, namely $\mathbf{V} = \mathbf{0}$ and $\mathbf{W}_\Pi = \mathbf{0}$. The corresponding formulation can be expressed as

$$\text{(P2)} \quad (\hat{\mathbf{X}}, \hat{\mathbf{A}}) = \arg\min_{(\mathbf{X},\mathbf{A})} \ \|\mathbf{X}\|_* + \lambda\|\mathbf{A}\|_1$$
$$\text{s.to} \quad \mathbf{Y} = \mathbf{R}\left(\mathbf{X} + \mathbf{A}\right), \quad \mathbf{Z}_\Pi = \mathscr{P}_\Pi\left(\mathbf{X} + \mathbf{A}\right).$$

In the sequel, identifiability of $(\mathbf{X}, \mathbf{A})$ from the linear measurements $\{\mathbf{Y}, \mathbf{Z}_\Pi\}$ is pursued first, followed by technical conditions based on certain incoherence measures, to guarantee $(\hat{\mathbf{X}} = \mathbf{X}_0, \hat{\mathbf{A}} = \mathbf{A}_0)$, where $\mathbf{X}_0$ and $\mathbf{A}_0$ are the *true* low-rank and sparse matrices of interest.

A. Local Identifiability

Let $r := \mathrm{rank}(\mathbf{X}_0)$ and $s := \|\mathbf{A}_0\|_0$ denote the rank and sparsity level of the true matrices of interest. The first issue to address is identifiability, asserting that there is a *unique* pair $(\mathbf{X}_0, \mathbf{A}_0)$ fulfilling the data constraints: (d1) $\mathbf{Y} = \mathbf{R}(\mathbf{X}_0 + \mathbf{A}_0)$ and (d2) $\mathbf{Z}_\Pi = \mathscr{P}_\Pi(\mathbf{X}_0 + \mathbf{A}_0)$. Apparently, if multiple solutions exist, one cannot hope finding $(\mathbf{X}_0, \mathbf{A}_0)$. Before examining this issue, introduce the subspaces: (s1) $\mathscr{N}_R := \{\mathbf{H} : \mathbf{RH} = \mathbf{0}_{L\times T}\}$ as the nullspace of the linear operator $\mathbf{R}$, and (s2) $\mathscr{N}_\Pi := \{\mathbf{H} \in \mathbb{R}^{F\times T} : \mathrm{supp}(\mathbf{H}) \subseteq \Pi^\perp\}$ as the nullspace of the linear operator $\mathscr{P}_\Pi(.)$ [$\Pi^\perp$ is the complement of Π]. If there exists a perturbation pair $(\mathbf{H}_1, \mathbf{H}_2)$ with $\mathbf{H}_1 + \mathbf{H}_2 \in \mathscr{N}_R \cap \mathscr{N}_\Pi$ so that $\mathbf{X}_0 + \mathbf{H}_1$ and $\mathbf{A}_0 + \mathbf{H}_2$ are still low-rank and sparse, one may pick the pair $(\mathbf{X}_0 + \mathbf{H}_1, \mathbf{A}_0 + \mathbf{H}_2)$ as another legitimate solution. This section aims at resolving such identifiability issues.

Let $\mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0'$ denote the singular value decomposition (SVD) of $\mathbf{X}_0$, and consider the subspaces: (s3) $\Phi_{X_0} := \{\mathbf{Z} \in \mathbb{R}^{F\times T} : \mathbf{Z} = \mathbf{U}_0\mathbf{W}_1' + \mathbf{W}_2\mathbf{V}_0',\ \mathbf{W}_1 \in \mathbb{R}^{T\times r},\ \mathbf{W}_2 \in \mathbb{R}^{F\times r}\}$ of matrices in either the column or row space of $\mathbf{X}_0$; (s4) $\Omega_{A_0} := \{\mathbf{H} \in \mathbb{R}^{F\times T} : \mathrm{supp}(\mathbf{H}) \subseteq \mathrm{supp}(\mathbf{A}_0)\}$ of matrices whose support is contained in that of $\mathbf{A}_0$. Noteworthy properties of these subspaces are: (i) since Φ_{X_0} and $\Omega_{A_0} \subset \mathbb{R}^{F\times T}$, it is possible to directly compare elements from them; (ii) $\mathbf{X}_0 \in \Phi_{X_0}$ and $\mathbf{A}_0 \in \Omega_{A_0}$; and (iii) if $\mathbf{Z} \in \Phi_{X_0}^\perp$ is added to $\mathbf{X}_0$, then $\mathrm{rank}(\mathbf{Z} + \mathbf{X}_0) > r$, and likewise $\mathbf{Z} \in \mathscr{N}_\Omega$, for any $\mathbf{Z} \in \Omega_{A_0}^\perp$.

Suppose temporarily that the subspaces Φ_{X_0} and Ω_{A_0} are also known. This extra piece of information helps identifiability based on data (d1) and (d2) since the potentially troublesome solutions

$$\Upsilon_1 := \{(\mathbf{X}_0 + \mathbf{H}_1, \mathbf{A}_0 + \mathbf{H}_2) : \mathbf{H}_1 + \mathbf{H}_2 \in \mathscr{N}_R \cap \mathscr{N}_\Pi\} \quad (5)$$

are restricted to a smaller set. If $(\mathbf{X}_0 + \mathbf{H}_1, \mathbf{A}_0 + \mathbf{H}_2) \notin \Upsilon_2$, where

$$\Upsilon_2 := \{(\mathbf{X}_0 + \mathbf{H}_1, \mathbf{A}_0 + \mathbf{H}_2) : \mathbf{H}_1 \in \Phi_{X_0},\ \mathbf{H}_2 \in \Omega_{A_0}\} \quad (6)$$

that candidate solution is not admissible since it is known a priori that $\mathbf{X}_0 \in \Phi_{X_0}$ and $\mathbf{A}_0 \in \Omega_{A_0}$. This notion of exploiting additional knowledge to assure uniqueness is known as *local identifiability* [14]. Global identifiability from (d1) and (d2) is not guaranteed. However, local identifiability will become essential later on to establish the main result. With these preliminaries, the following lemma puts forth the necessary and sufficient conditions for local identifiability.

Lemma 2: *Matrices $(\mathbf{X}_0, \mathbf{A}_0)$ satisfy (d1) and (d2) uniquely if and only if (c1) $\Phi_{X_0} \cap \Omega_{A_0} = \{\mathbf{0}\}$; and, (c2) $\Upsilon_1 \cap \Upsilon_2 = \{\mathbf{0}\}$.*

Condition (c1) implies that for the solutions in Υ_2 to be admissible, $\mathbf{H}_1 + \mathbf{H}_2$ must belong to the subspace $\Phi_{X_0} \oplus \Omega_{A_0}$. Accordingly, (c2) holds true if

$$\mathscr{N}_R \cap \mathscr{N}_\Pi \cap (\Phi_{X_0} \oplus \Omega_{A_0}) = \{\mathbf{0}\}. \quad (7)$$

Notice that (c1) appears also in the context of low-rank-plus-sparse recovery results in [14], [25]. However, (c2) is unique to the setting here. It captures the impact of the overlap between the nullspace of $\mathbf{R}$ and the operator $\mathscr{P}_\Pi(\cdot)$. Finding simpler sufficient conditions to assure (c1) and (c2) is dealt with next.

B. Incoherence Measures

The overlap between any pair of subspaces $\{\Phi_{X_0}, \Omega_{A_0}, \mathscr{N}_R, \mathscr{N}_\Pi\}$ plays a crucial role in identifiability and exact recovery as seen e.g., from Lemma 1. To quantify the overlap of the subspaces e.g., Φ_{X_0} and Ω_{A_0}, consider the *incoherence* parameter

$$\mu(\Phi_{X_0}, \Omega_{A_0}) := \max_{\substack{\mathbf{X}\in\Omega_{A_0} \\ \|\mathbf{X}\|_F=1}} \|\mathscr{P}_{\Phi_{X_0}}(\mathbf{X})\|_F, \quad (8)$$

which clearly satisfies $\mu(\Phi_{X_0}, \Omega_{A_0}) \in [0, 1]$. The lower bound is achieved when Φ_{X_0} and Ω_{A_0} are orthogonal, whereas the upperbound is attained when $\Phi_{X_0} \cap \Omega_{A_0}$ contains a nonzero element. To gain further geometric intuition, $\mu(\Phi_{X_0}, \Omega_{A_0})$ represents the cosine of the angle between subspaces when they have trivial intersection, namely $\Phi_{X_0} \cap \Omega_{A_0} = \{\mathbf{0}\}$ [26]. Small values of $\mu(\Phi_{X_0}, \Omega_{A_0})$ indicate sufficient separation between Φ_{X_0} and Ω_{A_0}, and thus less chance of ambiguity when discerning $\mathbf{X}_0$ from $\mathbf{A}_0$.

It will be seen later that (c1) requires $\mu(\Phi_{X_0}, \Omega_{A_0}) < 1$. In addition, to ensure (c2) one needs the incoherence parameter $\mu(\mathscr{N}_R \cap \mathscr{N}_\Pi, \Phi_{X_0} \oplus \Omega_{A_0}) < 1$. In fact, $\mu(\mathscr{N}_R \cap \mathscr{N}_\Pi, \Phi_{X_0} \oplus \Omega_{A_0})$ captures the ambiguity inherent to the nullspace of the compression and sampling operators. It depends on all subspaces (s1)–(s4), and it is desirable to express it in terms of the incoherence of different subspace pairs, namely $\mu(\mathscr{N}_R, \Omega_{A_0})$, $\mu(\mathscr{N}_R, \Phi_{X_0})$, $\mu(\mathscr{N}_\Pi, \Omega_{A_0})$, and $\mu(\mathscr{N}_\Pi, \Phi_{X_0})$. This is formalized in the next claim.

Proposition 1: *Assume that* $\mu(\Omega_{A_0},\Phi_{X_0}) < 1$. *If either* $\dim(\mathscr{N}_R\cap\mathscr{N}_\Pi)=0$*; or,* $\dim(\mathscr{N}_R\cap\mathscr{N}_\Pi)\geq 1$ *and*

$$\chi := \Big[\frac{\mu(\mathscr{N}_\Pi,\Phi_{X_0})+\mu(\mathscr{N}_R,\Omega_{A_0})\mu(\mathscr{N}_\Pi,\Omega_{A_0})}{1-\mu(\Omega_{A_0},\Phi_{X_0})}\Big]^{1/2} < 1$$

hold, then $\Phi_{X_0}\cap\Omega_{A_0}=\{\mathbf{0}\}$ *and* $\mathscr{N}_R\cap\mathscr{N}_\Pi\cap(\Phi_{X_0}\oplus\Omega_{A_0})=\{\mathbf{0}\}$.

Proof: Since $\mu(\Omega_{A_0},\Phi_{X_0})<1$ and $\dim(\Phi_{\mathrm{X_0}}\oplus\Omega_{\mathrm{A_0}}\oplus(\mathscr{N}_\mathrm{R}\cap\mathscr{N}_\Pi))=\dim(\Phi_{\mathrm{X_0}})+\dim(\Omega_{\mathrm{A_0}})+\dim(\mathscr{N}_\mathrm{R}\cap\mathscr{N}_\Omega)$, [22, Lemma 11] implies that

$$\mu^2(\Phi_{X_0}\oplus\Omega_{A_0},\mathscr{N}_R\cap\mathscr{N}_\Pi)\leq\big[1-\mu(\Phi_{X_0},\Omega_{A_0})\big]^{-1}\times\big[\mu^2(\Phi_{X_0},\mathscr{N}_R\cap\mathscr{N}_\Pi)+\mu^2(\Omega_{A_0},\mathscr{N}_R\cap\mathscr{N}_\Pi)\big]. \quad (9)$$

The result then follows by bounding $\mu^2(\Phi_{X_0},\mathscr{N}_R\cap\mathscr{N}_\Pi)\leq\mu(\Phi_{X_0},\mathscr{N}_R)\mu(\Phi_{X_0},\mathscr{N}_\Pi)$ using the fact that $\mathscr{N}_R\cap\mathscr{N}_\Pi\in\mathscr{N}_R,\mathscr{N}_\Pi$ [likewise for $\mu(\Omega_{A_0},\mathscr{N}_R\cap\mathscr{N}_\Pi)$], and $\mathscr{N}_R\cap\mathscr{N}_{\Phi_{X_0}}\neq\{\mathbf{0}\}$. ∎

Apparently, small values of $\mu(\mathscr{N}_R,\Omega_{A_0})$ and $\mu(\mathscr{N}_\Pi,\Phi_{X_0})$ gives rise to a small χ. In fact, $\mu(\mathscr{N}_R,\Omega_{A_0})$ measures whether $\mathscr{N}_R$ contains sparse elements, and it is tightly related to the incoherence among the sparse column-subsets of $\mathbf{R}$. For row-orthonormal compression matrices in particular, where $\mathbf{R}\mathbf{R}'=\mathbf{I}$, the incoherence reduces to the restricted isometry constant of $\mathbf{R}$, see e.g., [20]. Moreover, $\mu(\mathscr{N}_\Pi,\Phi_{X_0})$ measures whether the low-rank matrices fall into the nullspace of the subsampling operator $\mathscr{P}_\Pi(\cdot)$, that is tightly linked to the incoherence metrics introduced in the context of matrix completion; see e.g., [27]. It is worth mentioning that a wide class of matrices resulting in small incoherence $\mu(\mathscr{N}_R,\Omega_{A_0})$, $\mu(\mathscr{N}_\Pi,\Phi_{X_0})$ and $\mu(\Omega_{A_0},\Phi_{X_0})$ are provided in [20], [27], [25], which give rise to a sufficiently small value of χ.

C. Exact Recovery via Convex Optimization

Besides $\mu(\Omega_{A_0},\Phi_{X_0})$ and χ, there are other incoherence measures which play an important role in the conditions for exact recovery. These measures are introduced to avoid ambiguity when the (feasible) perturbations $\mathbf{H}_1$ and $\mathbf{H}_2$ do not necessarily belong to the subspaces Φ_{X_0} and Ω_{A_0}, respectively. Before moving on, it is worth noting that these measures resemble the ones for matrix completion and decomposition problems; see e.g., [25], [27]. For instance, consider a feasible solution $\{\mathbf{X}_0+a_{i,j}\mathbf{e}_i\mathbf{e}_j',\mathbf{A}_0+a_{i,j}\mathbf{e}_i\mathbf{e}_j'\}$, where $(i,j)\notin\mathrm{supp}(\mathbf{A}_0)$, and thus $a_{i,j}\mathbf{e}_i\mathbf{e}_j'\notin\Omega_{A_0}$. It may happen that $a_{i,j}\mathbf{e}_i\mathbf{e}_j'\in\Phi_{X_0}$ and $\mathrm{rank}(\mathbf{X}_0+a_{i,j}\mathbf{e}_i\mathbf{e}_j')=\mathrm{rank}(\mathbf{X}_0)-1$, while $\|\mathbf{A}_0-a_{i,j}\mathbf{e}_i\mathbf{e}_j'\|_0=\|\mathbf{A}_0\|_0+1$, thus challenging identifiability when Φ_{X_0} and Ω_{A_0} are unknown. Similar complications arise if $\mathbf{X}_0$ has a sparse row space that can be confused with the row space of $\mathbf{A}_0$. These issues motivate defining

$$\gamma(\mathbf{U}_0):=\max_i\|\mathbf{P}_U\mathbf{e}_i\|,\qquad\gamma(\mathbf{V}_0):=\max_i\|\mathbf{P}_V\mathbf{e}_i\| \quad (10)$$

where $\mathbf{P}_U:=\mathbf{U}_0\mathbf{U}_0'$ (resp. $\mathbf{P}_V:=\mathbf{V}_0\mathbf{V}_0'$) are the projectors onto the column (row) space of $\mathbf{X}_0$. Notice that $\gamma(\mathbf{U}_0),\gamma(\mathbf{V}_0)\in[0,1]$. The maximum of $\gamma(\mathbf{U}_0)$ (resp. $\gamma(\mathbf{V}_0)$) is attained when $\mathbf{e}_i$ is in the column (row) space of $\mathbf{X}_0$ for some i. Small values of $\gamma(\mathbf{U}_0)$ (resp. $\gamma(\mathbf{V}_0)$) imply that the column (row) spaces of $\mathbf{X}_0$ do not contain sparse vectors, respectively.

Another identifiability instance arises when $\mathbf{X}_0$ is sparse, in which case each column of $\mathbf{X}_0$ is spanned by a few canonical basis vectors. Consider the parameter

$$\gamma(\mathbf{U}_0,\mathbf{V}_0):=\|\mathbf{U}_0\mathbf{V}_0'\|_\infty=\max_{i,j}|\mathbf{e}_i'\mathbf{U}_0\mathbf{V}_0\mathbf{e}_j|. \quad (11)$$

A small value of $\gamma(\mathbf{U}_0,\mathbf{V}_0)$ indicates that each column of $\mathbf{X}_0$ is spanned by sufficiently many canonical basis vectors. It is worth noting that $\gamma(\mathbf{U}_0,\mathbf{V}_0)$ can be bounded in terms of $\gamma(\mathbf{U}_0)$ and $\gamma(\mathbf{V}_0)$, but it is kept here for the sake of generality.

From (c2) in Lemma 1 it is evident that the dimension of the nullspace $\mathscr{N}_R\cap\mathscr{N}_\Pi$ is critical for identifiability. In essence, the lower $\dim(\mathscr{N}_R\cap\mathscr{N}_\Pi)$ is, the higher is the chance for exact reconstruction. In order to quantify the size of the nullspace, define

$$\tau(\mathscr{N}_R,\mathscr{N}_\Pi):=\max_{\substack{\mathbf{X}\in\mathscr{N}_R\cap\mathscr{N}_\Pi\\\|\mathbf{X}\|=1}}\|\mathbf{X}\|_\infty \quad (12)$$

which will appear later in the exact recovery conditions. All elements are now in place to state the main result.

D. Main Result

Theorem 1: *Let* $(\mathbf{X}_0,\mathbf{A}_0)$ *denote the true low-rank and sparse matrix pair of interest, and define* $\mathbf{X}_0:=\mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0'$, $r:=$ *rank*$(\mathbf{X}_0)$, *and* $s:=\|\mathbf{A}_0\|_0$. *Assume that* $\mathbf{A}_0$ *has at most* k *nonzero elements per column, and define the incoherence parameters* $\alpha:=\mu(\Omega_{A_0},\Phi_{X_0})$, $\beta:=\mu(\Omega_{A_0},\mathscr{N}_R)$, $\xi:=\mu(\mathscr{N}_\Pi,\Phi_{X_0})$, $\nu:=\mu(\mathscr{N}_R,\Omega_{A_0}\cap\mathscr{N}_\Pi)$, $\eta:=\gamma(\mathbf{U}_0)+\gamma(\mathbf{V}_0)$, $\tau:=\tau(\mathscr{N}_R,\mathscr{N}_\Pi)$, $\gamma:=\gamma(\mathbf{U}_0,\mathbf{V}_0)$. *Given* $\mathbf{Y}$ *and* $\mathbf{Z}_\Pi$ *adhering to (d1) and (d2), respectively, with known* $\mathbf{R}$ *and* Π, *if* $\chi<1$, *and*

$$\text{(I)}\quad \lambda_{\max}:=(\frac{1}{k})\frac{1-\alpha-\alpha^3(1-\alpha^2)-ge/f}{1+\alpha^2(1-\alpha^2)+he/f} > \lambda_{\min}:=\frac{\gamma+qg/f}{1-\eta\alpha k-kqh/f}\geq 0$$

$$\text{(II)}\quad f:=1-\nu\beta-(\xi+\alpha\nu)(1-\alpha^2)(\xi+\alpha\beta)>0$$

hold, where

$$g:=\xi+\alpha(\xi+\alpha\nu)(1-\alpha^2)\alpha,\quad h:=\nu+\alpha(1-\alpha^2)(\xi+\alpha\nu)$$
$$q:=\tau+\eta\alpha+\eta\xi,\quad e:=\alpha(1-\alpha^2)(\xi+\alpha\beta)+1+\nu$$

then for any $\lambda_{\min}\leq\lambda\leq\lambda_{\max}$ *the convex program (P1) yields* $(\hat{\mathbf{X}}=\mathbf{X}_0,\hat{\mathbf{A}}=\mathbf{A}_0)$.

Satisfaction of the conditions in Theorem 1 hinges upon the incoherence parameters $\{\alpha,\gamma,\eta,\xi,\tau\}$ whose sufficiently small values fulfil (I) and (II). In fact, these parameters are increasing functions of the rank r and the sparsity level s. In particular, $\{\alpha,\gamma,\eta\}$ that capture the ambiguity of the additive components $\mathbf{X}_0$ and $\mathbf{A}_0$, are known to be small enough for small values of $\{r,s,k\}$; see e.g., [14], [27]. Regarding χ, recall that it is an increasing function of β and ξ, where the parameter ξ takes a small value when NetFlow samples an adequately large subset of OD flows uniformly at random. Moreover, in large-scale networks with distant OD node pairs, and routing paths that are sufficiently "spread-out", the sparse

column-subsets of $\mathbf{R}$ tend to be incoherent, and thus β takes a small value. Likewise, for sufficiently many NetFlow samples and column-incoherent routing matrices, τ takes a small value.

Remark 1 (Satisfiability): Notice that (I) and (II) in Theorem 1 are expressible in terms of the angle between subspaces (s1)–(s4). In general, they are NP-hard to verify. Introducing a class of (possibly random) traffic matrices $(\mathbf{X}_0, \mathbf{A}_0)$ and realistic network settings giving rise to a desirable routing matrix $\mathbf{R}$ is the subject of our ongoing research. The major roadblock in this direction is deriving tight bounds for the parameter τ, which involves the intersection of a pair of subspaces.

E. ADMM Algorithm

This section introduces an iterative solver for the convex program (P2) using the alternating direction method of multipliers (ADMM) method. ADMM is an iterative augmented Lagrangian method especially well-suited for parallel processing [28], and has been proven successful to tackle the optimization tasks encountered e.g., in statistical learning; see e.g., [29]. While ADMM could be directly applied to (P2), $\mathbf{R}$ couples the entries of $\mathbf{A}$ and $\mathbf{X}$ leading to computationally demanding nuclear- and ℓ_1-norm minimization subtasks per iteration. To overcome this hurdle, a common trick is to introduce auxiliary (decoupling) variables $\{\mathbf{B}, \mathbf{O}\}$, and formulate the following optimization problem

$$\begin{aligned}(\text{P3}) \quad \min_{\{\mathbf{A},\mathbf{X},\mathbf{O},\mathbf{B}\}} & \|\mathbf{X}\|_* + \lambda\|\mathbf{A}\|_1 \\ \text{s. to } & \mathbf{Y} = \mathbf{R}(\mathbf{O}+\mathbf{B}), \quad \mathbf{Z}_\Pi = \mathscr{P}_\Pi(\mathbf{O}+\mathbf{B}) \\ & \mathbf{B} = \mathbf{A}, \quad \mathbf{O} = \mathbf{X},\end{aligned}$$

which is equivalent to (P2). To tackle (P3), associate the Lagrange multipliers $\{\mathbf{M}_y, \mathbf{M}_z, \mathbf{M}_a, \mathbf{M}_x\}$ with the constraints, and then introduce the quadratically *augmented* Lagrangian function

$$\begin{aligned}&\mathscr{L}(\mathbf{X},\mathbf{A},\mathbf{B},\mathbf{O};\mathbf{M}_y,\mathbf{M}_z,\mathbf{M}_a,\mathbf{M}_x) \\ &:= \|\mathbf{X}\|_* + \lambda\|\mathbf{A}\|_1 + \langle \mathbf{M}_y, \mathbf{Y} - \mathbf{R}(\mathbf{O}+\mathbf{B})\rangle + \langle \mathbf{M}_a, \mathbf{B}-\mathbf{A}\rangle \\ &+ \langle \mathbf{M}_z, \mathbf{Z}_\Pi - \mathscr{P}_\Pi(\mathbf{O}+\mathbf{B})\rangle + \langle \mathbf{M}_x, \mathbf{O}-\mathbf{X}\rangle \\ &+ \frac{c}{2}\|\mathbf{Y} - \mathbf{R}(\mathbf{O}+\mathbf{B})\|_F^2 + \frac{c}{2}\|\mathbf{Z}_\Pi - \mathscr{P}_\Pi(\mathbf{O}+\mathbf{B})\|_F^2 \\ &+ \frac{c}{2}\|\mathbf{B}-\mathbf{A}\|_F^2 + \frac{c}{2}\|\mathbf{O}-\mathbf{X}\|_F^2 \qquad (13)\end{aligned}$$

where $c > 0$ is a penalty coefficient. Splitting the primal variables into two groups $\{\mathbf{X}, \mathbf{B}\}$ and $\{\mathbf{A}, \mathbf{O}\}$, the ADMM solver entails an iterative procedure comprising three steps per iteration $k = 1, 2, \ldots$

[S1] Update dual variables:

$$\mathbf{M}_y[k] = \mathbf{M}_y[k-1] + c(\mathbf{Y} - \mathbf{R}(\mathbf{O}[k]+\mathbf{B}[k])) \quad (14)$$
$$\mathbf{M}_z[k] = \mathbf{M}_z[k-1] + c(\mathbf{Z}_\Pi - \mathscr{P}_\Pi(\mathbf{O}+\mathbf{B})) \quad (15)$$
$$\mathbf{M}_a[k] = \mathbf{M}_a[k-1] + c(\mathbf{B}[k] - \mathbf{A}[k]) \quad (16)$$
$$\mathbf{M}_x[k] = \mathbf{M}_x[k-1] + c(\mathbf{O}[k] - \mathbf{X}[k]) \quad (17)$$

[S2] Update first group of primal variables:

$$\begin{aligned}&\mathbf{A}[k+1] \\ &= \arg\min_{\mathbf{A}\in\mathbb{R}^{F\times T}} \left\{\frac{c}{2}\|\mathbf{A} - \mathbf{B}[k]\|_F^2 - \langle \mathbf{M}_a[k], \mathbf{A}\rangle + \lambda\|\mathbf{A}\|_1\right\}. \\ &\mathbf{O}[k+1] \\ &= \arg\min_{\mathbf{O}\in\mathbb{R}^{F\times T}} \left\{\frac{c}{2}\|\mathbf{O} - \mathbf{X}[k]\|_F^2 + \frac{c}{2}\|\mathbf{Y} - \mathbf{R}(\mathbf{O}+\mathbf{B}[k])\|_F^2 \right. \\ &\qquad + \frac{c}{2}\|\mathbf{Z}_\Pi - \mathscr{P}_\Pi(\mathbf{O}+\mathbf{B}[k])\|_F^2 \\ &\qquad \left. + \langle \mathbf{M}_x[k] - \mathbf{R}'\mathbf{M}_y[k] - \mathscr{P}_\Pi(\mathbf{M}_z[k]), \mathbf{O}\rangle\right\}.\end{aligned}$$

[S3] Update second group of primal variables:

$$\begin{aligned}&\mathbf{X}[k+1] \\ &= \arg\min_{\mathbf{X}\in\mathbb{R}^{F\times T}} \left\{\frac{c}{2}\|\mathbf{X} - \mathbf{O}[k]\|_F^2 - \langle \mathbf{M}_x[k], \mathbf{X}\rangle + \|\mathbf{X}\|_*\right\} \\ &\mathbf{B}[k+1] \\ &= \arg\min_{\mathbf{B}\in\mathbb{R}^{F\times T}} \left\{\frac{c}{2}\|\mathbf{A}[k] - \mathbf{B}\|_F^2 + \frac{c}{2}\|\mathbf{Y} - \mathbf{R}(\mathbf{O}[k]+\mathbf{B})\|_F^2 \right. \\ &\qquad + \frac{c}{2}\|\mathbf{Z}_\Pi - \mathscr{P}_\Pi(\mathbf{O}[k]+\mathbf{B})\|_F^2 \\ &\qquad \left. + \langle \mathbf{M}_a[k] - \mathbf{R}'\mathbf{M}_y[k] - \mathscr{P}_\Pi(\mathbf{M}_z[k]), \mathbf{B}\rangle\right\}\end{aligned}$$

The resulting iterative solver is tabulated under Algorithm 1. Here, $[\mathscr{S}_\tau(\mathbf{X})]_{i,j} := \text{sgn}(x_{i,j})\max\{|x_{i,j}| - \tau, 0\}$ refers to the soft-thresholding operator; the vectors $\{\mathbf{y}_t, \mathbf{o}_t, \mathbf{a}_t, \mathbf{b}_t, \mathbf{z}_t, \mathbf{x}_t, \mathbf{m}_t^z, \mathbf{m}_t^a, \mathbf{m}_t^x, \mathbf{m}_t^y\}$ denote the t-th column of their corresponding matrix arguments, and the diagonal matrix $\mathbf{\Pi}_t \in \{0,1\}^{P\times P}$ is unity at (i,i)-th entry if $(i,t) \in \Pi$, and zero otherwise. Algorithm 1 reveals that the update for the anomaly matrix entails a soft-thresholding operator to promote sparsity, while the nominal traffic is updated via singular value thresholding to effect low rank. The updates for $\mathbf{B}$ and $\mathbf{O}$ are also parallelized across the rows. Due to convexity of (P3), Algorithm 1 with two Gauss-Seidel block updates is convergent to the global optimum of (P2) as stated next.

Proposition 2: *[28] For any value of the penalty coefficient $c > 0$, the iterates $\{\mathbf{X}[k], \mathbf{A}[k]\}$ converge to the optimal solution of (P2) as $k \to \infty$.*

V. Incorporating Spatiotemporal Correlation Information

Being convex (P1) is appealing, and as Theorem 1 asserts for the noiseless case it reconstructs reliably the underlying traffic when: (c1) the anomalous traffic is sufficiently "sporadic" across time and flows; (c2) the nominal traffic matrix is sufficiently low-rank with non-spiky singular vectors; (c3) NetFlow *uniformly* samples OD flows; and, (c4) the routing paths are sufficiently "spread-out." In practical networks however, these conditions may be violated, and as a consequence (P1) may perform poorly. For instance, if a bursty anomaly occurs, (c1) does not hold. A particular OD flow may also be inaccessible to sample via NetFlow, that violates (c3). Apparently, in the latter case, knowing the cross-correlation of a missing OD flow with other flows enables accurate interpolation of misses.

Inherent patterns of the nominal traffic matrix $\mathbf{X}$ and the anomalous traffic matrix $\mathbf{A}$ can be learned from historical/training data $\{\mathbf{x}_t, \mathbf{a}_t\}_{t\in\mathscr{H}}$, where $\mathbf{x}_t$ and $\mathbf{a}_t$ denote

Algorithm 1 : ADMM solver for (P2)

input $\mathbf{Y}, \mathbf{Z}_\Pi, \Pi, \mathbf{R}, \lambda, c, \{\mathbf{H}_t := (\mathbf{I}_F + \mathbf{\Pi}_t + \mathbf{R}'\mathbf{R})^{-1}\}_{t=1}^T$
initialize $\mathbf{M}_y[-1] = \mathbf{0}_{L\times T}$, $\mathbf{X}[0] = \mathbf{O}[0] = \mathbf{A}[0] = \mathbf{B}[0] = \mathbf{M}_z[-1] = \mathbf{M}_a[-1] = \mathbf{M}_x[-1] = \mathbf{0}_{F\times T}$, and set $k = 0$.
while not converged **do**
[S1] Update dual variables:
$\mathbf{M}_y[k] = \mathbf{M}_y[k-1] + c(\mathbf{Y} - \mathbf{R}(\mathbf{O}[k] + \mathbf{B}[k]))$
$\mathbf{M}_z[k] = \mathbf{M}_z[k-1] + c(\mathbf{Z}_\Pi - \mathscr{P}_\Pi(\mathbf{O}[k] + \mathbf{B}[k]))$
$\mathbf{M}_a[k] = \mathbf{M}_a[k-1] + c(\mathbf{B}[k] - \mathbf{A}[k])$
$\mathbf{M}_x[k] = \mathbf{M}_x[k-1] + c(\mathbf{O}[k] - \mathbf{X}[k])$
[S2] Update first group of primal variables:
$\mathbf{A}[k+1] = \mathscr{S}_{\frac{1}{c}}(c^{-1}\mathbf{M}_a[k] + \mathbf{B}[k])$.
Update in parallel ($t = 1, \ldots, T$)
$\mathbf{o}_t[k+1] = \mathbf{H}_t\big(c\mathbf{x}_t[k] + c\mathbf{\Pi}_t\mathbf{z}_t + c\mathbf{R}'\mathbf{y}_t - c[\mathbf{\Pi}_t + \mathbf{R}'\mathbf{R}]\mathbf{b}_t[k] + \mathbf{R}'\mathbf{m}_t^y[k] + \mathbf{\Pi}_t\mathbf{m}_t^z[k] - \mathbf{m}_t^x[k]\big)$
[S3] Update second group of primal variables:
$\mathbf{U}\mathbf{\Sigma}\mathbf{V}' = \mathrm{svd}(\mathbf{O}[k+1] + c^{-1}\mathbf{M}_x[k])$, $\mathbf{X}[k+1] = \mathbf{U}\mathscr{S}_{1/c}(\mathbf{\Sigma})\mathbf{V}'$
Update in parallel ($t = 1, \ldots, T$)
$\mathbf{b}_t[k+1] = \mathbf{H}_t\big(c\mathbf{a}_t[k+1] + c\mathbf{\Pi}_t\mathbf{z}_t + c\mathbf{R}'\mathbf{y}_t - c[\mathbf{\Pi}_t + \mathbf{R}'\mathbf{R}]\mathbf{o}_t[k+1] + \mathbf{R}'\mathbf{m}_t^y[k] + \mathbf{\Pi}_t\mathbf{m}_t^z[k] - \mathbf{m}_t^a[k]\big)$
$k \leftarrow k+1$
end while
return $(\mathbf{A}[k], \mathbf{X}[k])$

the network-wide nominal and anomalous traffic vectors at time t. Given the training data $\{\mathbf{x}_t, \mathbf{a}_t\}_{t\in\mathscr{H}}$, link counts $\mathbf{Y}$ obeying (2) as well as the partial flow-counts $\mathbf{Z}_\Pi$ adhering to (4), and with $\{\mathbf{R}, \Pi\}$ known, the rest of this paper deals with estimating the matrix pair $(\mathbf{X}, \mathbf{A})$.

A. Bilinear Factorization

The first step toward incorporating correlation information is to use the bilinear characterization of the nuclear norm. Using singular value decomposition [24], one can always factorize the low-rank component as $\mathbf{X} = \mathbf{L}\mathbf{Q}'$, where $\mathbf{L} \in \mathbb{R}^{F\times\rho}$, $\mathbf{Q} \in \mathbb{R}^{T\times\rho}$, for some $\rho \geq \mathrm{rank}(\mathbf{X})$. The nuclear-norm can then be redefined as (see e.g., [30])

$$\|\mathbf{X}\|_* := \min_{\mathbf{X}=\mathbf{L}\mathbf{Q}'} \frac{1}{2}\{\|\mathbf{L}\|_F^2 + \|\mathbf{Q}\|_F^2\}. \tag{18}$$

For the scalar case, (18) leads to the identity $|a| = \min_{a=bc}\frac{1}{2}(|b|^2 + |c|^2)$. The latter implies that the ℓ_1-norm of $\mathbf{A}$ can be alternatively defined as

$$\|\mathbf{A}\|_1 := \min_{\mathbf{A}=\mathbf{B}\odot\mathbf{C}} \frac{1}{2}\{\|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2\} \tag{19}$$

where $\mathbf{B}, \mathbf{C} \in \mathbb{R}^{F\times T}$. For notational convenience, let $\mathbf{U} := [\mathbf{Y}', \ \mathbf{Z}_\Pi']$ and the corresponding linear operator $\mathscr{P}(\mathbf{X}) := [(\mathbf{R}\mathbf{X})', \ \mathscr{P}_\Omega(\mathbf{X})']$. Leveraging (18) and (19), one is prompted to recast (P1) as

$$\text{(P4)} \quad \min_{\{\mathbf{L},\mathbf{Q},\mathbf{B},\mathbf{C}\}} \ \frac{1}{2}\|\mathbf{U} - \mathscr{P}(\mathbf{L}\mathbf{Q}' + \mathbf{B}\odot\mathbf{C})\|_F^2 + \frac{\lambda_*}{2}\big\{\|\mathbf{L}\|_F^2 + \|\mathbf{Q}\|_F^2\big\} + \frac{\lambda_1}{2}\big\{\|\mathbf{B}\|_F^2 + \|\mathbf{C}\|_F^2\big\}.$$

This Frobenius-norm regularization doubles the number of optimization variables for the sparse component $\mathbf{A}$ ($2FT$), but reduces the variable count for the low-rank component $\mathbf{X}$ to $\rho(F+T)$. Regarding performance, the bilinear factorization incurs no loss of optimality as stated in the next lemma.

Lemma 3: *If $\hat{\mathbf{X}}$ denotes the optimal low-rank solution of (P1) and $\rho \geq \mathrm{rank}(\hat{\mathbf{X}})$, then (P4) is equivalent to (P1).*

Proof: It readily follows from (18) and (19) along with the commutative property of minimization which allows taking minimization first with respect to (w.r.t.) $\{\mathbf{L}, \mathbf{Q}\}$ and then w.r.t. $\{\mathbf{B}, \mathbf{C}\}$. ■

VI. Bayesian Traffic and Anomaly Estimates

This section recasts (P4) in a Bayesian framework by adopting the AWGN model $\mathbf{U} = \mathscr{P}(\mathbf{X} + \mathbf{A}) + \mathbf{E}$, where $\mathbf{E}$ contains independent identically distributed (i.i.d.) entries drawn from $\mathscr{N}(0, \sigma^2)$. As in (18) $\mathbf{X}$ is also factorized as $\mathbf{L}\mathbf{Q}'$ with the independent factors $\mathbf{L} := [\mathbf{l}_1, \ldots, \mathbf{l}_\rho]$ and $\mathbf{Q} := [\mathbf{q}_1, \ldots, \mathbf{q}_\rho]$. Matrices $\mathbf{L}$ and $\mathbf{Q}$ are formed by i.i.d. columns obeying $\mathbf{l}_i \sim \mathscr{N}(0, \mathbf{R}_L)$ and $\mathbf{q}_i \sim \mathscr{N}(0, \mathbf{R}_Q)$, respectively, for positive-definite correlation matrices $\mathbf{R}_L \in \mathbb{R}^{F\times F}$ and $\mathbf{R}_Q \in \mathbb{R}^{T\times T}$. Without loss of generality (w.l.o.g.), in order to avoid the scalar ambiguity in $\mathbf{X} = \mathbf{L}\mathbf{Q}'$ set $\mathrm{tr}(\mathbf{R}_\mathrm{L}) = \mathrm{tr}(\mathbf{R}_\mathrm{Q})$. Likewise, the anomaly matrix is factored as $\mathbf{A} = \mathbf{B}\odot\mathbf{C}$ with the independent factors $\mathbf{b} := \mathrm{vec}(\mathbf{B}) \in \mathbb{R}^{FT}$ and $\mathbf{c} := \mathrm{vec}(\mathbf{C}) \in \mathbb{R}^{FT}$ drawn from $\mathbf{b} \sim \mathscr{N}(0, \mathbf{R}_B)$ and $\mathbf{c} \sim \mathscr{N}(0, \mathbf{R}_C)$, with positive-definite correlation matrices $\mathbf{R}_B, \mathbf{R}_C \in \mathbb{R}^{FT\times FT}$, respectively.

For the considered AWGN model with priors, the maximum a posteriori (MAP) estimator of $(\mathbf{X}, \mathbf{A})$ is given by the solution of

$$\text{(P5)} \quad \min_{\{\mathbf{L},\mathbf{Q},\mathbf{B},\mathbf{C}\}} \frac{1}{2}\|\mathbf{U} - \mathscr{P}(\mathbf{L}\mathbf{Q}' + \mathbf{B}\odot\mathbf{C})\|_F^2 + \frac{\lambda_1}{2}\big[\mathbf{b}'\mathbf{R}_B^{-1}\mathbf{b} + \mathbf{c}'\mathbf{R}_C^{-1}\mathbf{c}\big] + \frac{\lambda_*}{2}\big[\mathrm{tr}(\mathbf{L}'\mathbf{R}_L^{-1}\mathbf{L}) + \mathrm{tr}(\mathbf{Q}'\mathbf{R}_Q^{-1}\mathbf{Q})\big]$$

for $\lambda_1 = \lambda_* = \sigma^2$, where different weights λ_1 and λ_* are considered here for generality. Observe that (P5) specializes to (P4) upon choosing $\mathbf{R}_L = \mathbf{I}_F$, $\mathbf{R}_Q = \mathbf{I}_T$, and $\mathbf{R}_B = \mathbf{R}_C = \mathbf{I}_{FT}$. Lemma 3 then implies that the convex program (P1) yields the MAP optimal estimator for the considered statistical model so long as the factors contain i.i.d. Gaussian entries. With respect to the statistical model for the low-rank and sparse components, as it will become clear later on, $\mathbf{R}_L$ ($\mathbf{R}_Q$) captures the correlation among columns (rows) of $\mathbf{X}$; likewise, $\mathbf{R}_B$ and $\mathbf{R}_C$ capture the correlation among entries of $\mathbf{A}$.

Albeit clear in this section statistical formulation, the adopted model $\mathbf{X} = \mathbf{L}\mathbf{Q}'$ promotes low rank as a result of $\mathrm{rank}(\mathbf{X}) \leq \rho$, but it is not obvious whether $\mathbf{A} = \mathbf{B}\odot\mathbf{C}$ effects sparsity. The latter will rely on the fact that the product of two independent Gaussian random variables is heavy tailed. To recognize this, consider the independent scalar random variables $b \sim \mathscr{N}(0,1)$ and $c \sim \mathscr{N}(0,1)$. The product random variable $a = bc$ can then be expressed as $bc = \frac{1}{4}(b+c)^2 - \frac{1}{4}(b-c)^2$, where $S_1 := \frac{1}{4}(b+c)^2$ and $S_2 := \frac{1}{4}(b-c)^2$ are central χ^2-distributed random variables. Since $\mathbb{E}[(a-b)(a+b)] = 0$, the random variables S_1 and S_2 are independent, and consequently the characteristic function of a admits the simple form $\Phi_a(\omega) = \Phi_{S_1}(\omega)\Phi_{S_2}(\omega) = 1/(\sqrt{1+4\omega^2})$. Applying the inverse Fourier transform to $\Phi_a(\omega)$, yields the probability density function $p_a(x) = (1/\sqrt{2\pi})k_0(x/2)$, where $k_0(x) := \int_0^\infty[\cos(\omega x)]/(\sqrt{1+4\omega^2})\, d\omega$ denotes the modified Bessel function of second-kind, which is tightly approximated

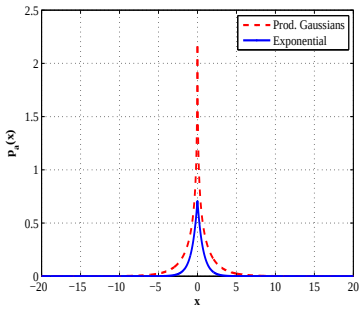


Fig. 2. Sparsity promoting priors with zero mean and unity variance.

with $\sqrt{\pi/(2x)}e^{-x}$ for $x > 1$ [31, p. 20]. One can then readily deduce that $p_a(x) = \sqrt{\pi/(2x)}e^{-|x|}$ behaves similar to the Laplacian distribution, which is well known to promote sparsity. In contrast with the Laplacian distribution however, the product of Gaussian random variables incurs a slightly lighter tail as depicted in Fig. 2. It is worth commenting that the correlated multivariate Laplacian distribution is an alternative prior distribution to postulate for the sparse component. However, its complicated form [32] renders the optimization for the MAP estimator intractable.

Remark 2 (nonzero mean): In general, one can allow nonzero mean for the factors in the adopted statistical model, and subsequently replaces correlations with covariances. This can be useful e.g., to estimate the nominal traffic which is inherently positive valued. The mean values are assumed zero here for simplicity.

A. Learning the correlation matrices

Implementing (P5) requires first obtaining the correlation matrices $\{\mathbf{R}_L, \mathbf{R}_Q, \mathbf{R}_B, \mathbf{R}_C\}$ from the second-order statistics of $(\mathbf{X}, \mathbf{A})$, or their estimates based on training data. Given second-order statistics of the unknown nominal-traffic matrix $\mathbf{X}$, matrices $\{\mathbf{R}_L, \mathbf{R}_Q\}$ can be readily found as explained in the next lemma. The proof is along the lines of [33], hence it is omitted for brevity.

Lemma 4: *Under the Gaussian bilinear model for* $\mathbf{X}$*, and with* $\mathrm{tr}(\mathbf{R}_\mathrm{L}) = \mathrm{tr}(\mathbf{R}_\mathrm{Q})$*, it holds that*

$$\mathbf{R}_Q = \rho\mathbb{E}[\mathbf{X}'\mathbf{X}]/(\mathbb{E}[\|\mathbf{X}\|_F^2])^{1/2},$$
$$\mathbf{R}_L = \rho\mathbb{E}[\mathbf{X}\mathbf{X}']/(\mathbb{E}[\|\mathbf{X}\|_F^2])^{1/2}.$$

It is evident that $\mathbf{R}_L$ captures *temporal* correlation of the network traffic (columns of $\mathbf{X}$), while $\mathbf{R}_Q$ captures the *spatial* correlation across OD flows (rows of $\mathbf{X}$).

For real data where the distribution of unknowns is not available, $\{\mathbf{R}_L, \mathbf{R}_Q\}$ are typically estimated from the training data, which can be e.g., past estimates of nominal and anomalous traffic. For instance, consider $\{\mathbf{R}_L, \mathbf{R}_Q\}$ estimates as input to (P5) for estimating the traffic at day $K+1$ (corresponding to time horizon $\mathscr{T}$) with T time instants, from the training data $\{\mathbf{x}_t\}_{t=1}^{KT}$ collected during the past K days. Apparently, reliable correlation estimates cannot be formed for general nonstationary processes. Empirical analysis of Internet traffic suggests adopting the following assumptions [3]: (a1) Process $\{\mathbf{x}_t\}$ is cyclostationary with a day-long period due to large-scale periodic trends in the nominal traffic; and (a2) OD flows are uncorrelated as their origins are mutually unrelated. One can also take into account weekly or monthly periodicity of traffic usage to further improve the accuracy of the correlation estimates.

Let r_t denote the remainder of dividing t by T. For time slots $t_1, t_2 \in \mathscr{T}$, (a1) asserts that the vector subprocesses $\{\mathbf{x}_{kT+r_{t_1}}\}_{k=0}^{K-1}$ and $\{\mathbf{x}_{kT+r_{t_2}}\}_{k=0}^{K-1}$ are stationary, and thus one can consistently estimate $\mathbb{E}[\mathbf{x}'_{r_{t_1}}\mathbf{x}_{r_{t_2}}]$, to obtain $\mathbf{R}_Q$ via the sample correlation $\frac{1}{K}\sum_{k=0}^{K-1}\mathbf{x}_{kT+r_{t_1}}\mathbf{x}'_{kT+r_{t_2}}$ [34]. Likewise, the normalization term $\mathbb{E}[\|\mathbf{X}\|_F^2]$ is estimated relying on (a1) as $\frac{1}{K}\sum_{t=1}^{T}\sum_{k=0}^{K-1}\|\mathbf{x}_{kT+t}\|^2$. Estimating $\mathbf{R}_L$ on the other hand relies on (a2). Let $\boldsymbol{\xi}_f^{'} \in \mathbb{R}^T$ denote the time-series of traffic associated with OD flow f, namely the f-th row of $\mathbf{X}$. It then follows from (a2) that $\mathbb{E}[\boldsymbol{\xi}_{f_1}\boldsymbol{\xi}_{f_2}] = (\mathbb{E}[\boldsymbol{\xi}_{f_1}])'(\mathbb{E}[\boldsymbol{\xi}_{f_2}])$ for $f_1 \neq f_2 \in \mathscr{F}$, where due to (a1), $\mathbb{E}[\xi_{f,t}]$ ($\xi_{f,t}$ signifies the t-th entry of $\boldsymbol{\xi}_f$) is estimated via the sample mean $\frac{1}{K}\sum_{k=0}^{K-1}x_{f,kT+r_t}$. Moreover, for $f_1 = f_2 = f$, the estimate for $\mathbb{E}[\boldsymbol{\xi}'_f\boldsymbol{\xi}_f]$ is $\frac{1}{K}\sum_{k=0}^{K-1}\sum_{t=1}^{T}\xi^2_{f,kT+r_t}$.

Given the second-order statistics of $\mathbf{A}$, the correlation matrices $\mathbf{R}_B$ and $\mathbf{R}_C$ are obtained next.

Lemma 5: *Under the Gaussian bilinear model for* $\mathbf{a} = \mathrm{vec}(\mathbf{A}')$*, it holds that* $\mathbb{E}[\mathbf{a}\mathbf{a}'] = \mathbf{R}_B \odot \mathbf{R}_C$.

In order to avoid the scalar ambiguity present in $\mathbf{R}_B$ and $\mathbf{R}_C$, assume equal-magnitude entries $|[\mathbf{R}_B]_{i,j}| = |[\mathbf{R}_C]_{i,j}| = |\left[\mathbb{E}[\mathbf{a}\mathbf{a}']\right]_{i,j}|^{1/2}$, $\forall(i,j)$. Apparently, for a diagonal correlation matrix $\mathbb{E}[\mathbf{a}\mathbf{a}']$, the factors are uniquely determined as $[\mathbf{R}_B]_{i,i} = [\mathbf{R}_C]_{i,i} = \left[\mathbb{E}[\mathbf{a}\mathbf{a}']\right]_{i,i}^{1/2}$, $\forall i$. However, when nonzero off-diagonals are present, there may exist a sign ambiguity, and the signs should be assigned appropriately to guarantee that $\mathbf{R}_B$ and $\mathbf{R}_C$ are positive definite.

Correlation matrices $\{\mathbf{R}_B, \mathbf{R}_C\}$ required to run (P5) over the time horizon $\mathscr{T}$ ($|\mathscr{T}| = T$) are estimated from the training data $\{\mathbf{a}_t\}_{t=1}^{KT}$ collected e.g., over the past K days. Due to the diverse nature of anomalies, developing a universal methodology to learn $\mathbf{R}_B$ and $\mathbf{R}_C$ is an ambitious objective. Depending on the nature of anomalies, the learning process is possible under certain assumptions. One such reasonable assumption is that anomalies of different flows are uncorrelated, but for each OD flow, the anomalous traffic is stationary and possibly correlated over time. This model is appropriate e.g., when different flows are subject to bursty anomalies arising from unrelated external sources.

For the stationary anomaly process of flow f, namely $\{a_{f,t}\}_t$, let $R_a^{(f)}(\tau) := \mathbb{E}[a_{f,t-\tau}a_{f,t}]$ denote the time-invariant cross-correlation. Let also $\boldsymbol{\alpha}_f^{'}$ denote the f-th row of $\mathbf{A}$, and introduce the correlation matrix $\mathbf{R}_a^{(f)} := \mathbb{E}[\boldsymbol{\alpha}_f\boldsymbol{\alpha}_f^{'}] \in \mathbb{R}^T$, which is Toeplitz with entries $[\mathbf{R}_a^{(f)}]_{i,i+\tau} = R_a^{(f)}(\tau)$, $i \in [T], \tau = 0, \ldots, T-1$. Accordingly, $\mathbb{E}[\mathbf{a}\mathbf{a}']$ is a block-diagonal matrix with blocks $\mathbf{R}_a^{(f)}$, and subsequently Lemma 5 implies that $\mathbf{R}_B$ and $\mathbf{R}_C$ are block diagonal with Toeplitz blocks $\mathbf{R}_b^{(f)}$ and $\mathbf{R}_c^{(f)}$, respectively. Under the equal-magnitude assumption for the entries of $\mathbf{R}_B$ and $\mathbf{R}_c$, the entries of $\mathbf{R}_b^{(f)}$ and $\mathbf{R}_c^{(f)}$

are readily obtained as

$$\big[\mathbf{R}_b^{(f)}\big]_{i,i+\tau}=\big|R_a^{(f)}(\tau)\big|^{1/2},$$
$$\big[\mathbf{R}_c^{(f)}\big]_{i,i+\tau}=\big|R_a^{(f)}(\tau)\big|^{1/2}\mathrm{sgn}\big(R_a^{(f)}(\tau)\big). \quad (20)$$

Notice that if $|R_a^{(f)}(\tau)|$ decays sufficiently fast as τ grows, $\mathbf{R}_B$ and $\mathbf{R}_C$ become positive definite [35]. Finally, thanks to the stationarity of $\{a_{f,t}\}_t$, $R_a(\tau)$ can be consistently estimated using $\frac{1}{KT-\tau}\sum_{t=\tau+1}^{KT} a_{f,t-\tau}a_{f,t}$. It is worth noting that the considered model renders the sparsity regularizer in (P5) separable across rows of $\mathbf{A}$, which in turn induces row-wise sparsity.

VII. Alternating Majorization-Minimization Algorithm

In order to efficiently solve (P5), an alternating minimization (AM) scheme is developed here by alternating among four matrix variables $\{\mathbf{L},\mathbf{Q},\mathbf{B},\mathbf{C}\}$. The algorithm entails iterations updating one matrix variable at a time, while keeping the rest are kept fixed at their up-to-date values. In particular, iteration k comprises orderly updates of four matrices $\mathbf{L}[k]\rightarrow\mathbf{Q}[k]\rightarrow\mathbf{B}[k]\rightarrow\mathbf{C}[k]$. For instance, $\mathbf{L}[k]$ is updated given the latest updates $\{\mathbf{Q}[k-1],\mathbf{B}[k-1],\mathbf{C}[k-1]\}$ as $\mathbf{L}[k]=\arg\min_{\mathbf{L}} g_L^{(k)}(\mathbf{L})$, where

$$g_L^{(k)}(\mathbf{L}):=\frac{1}{2}\|\mathbf{U}-\mathscr{P}(\mathbf{L}\mathbf{Q}'[k-1]+\mathbf{B}[k-1]\odot\mathbf{C}[k-1])\|_F^2 + \frac{\lambda_*}{2}\mathrm{tr}\big(\mathbf{L}'\mathbf{R}_{\mathrm{L}}^{-1}\mathbf{L}\big) \quad (21)$$

Likewise, $\mathbf{Q}[k]$, $\mathbf{B}[k]$, and $\mathbf{C}[k]$ are updated by respectively minimizing $g_Q^{(k)}, g_B^{(k)}$, and $g_C^{(k)}$, which are given similar to $g_L^{(k)}$ based on latest updates of the corresponding variables.

Functions $\{g_L^{(k)},g_Q^{(k)},g_B^{(k)},g_C^{(k)}\}$ are strongly convex quadratic programs due to regularization with positive definite correlations in the regularizer, and thus their solutions admits closed form after inverting certain possibly large-size matrices. For instance, updating $\mathbf{L}[k]$ requires inverting an $F\rho\times F\rho$ matrix. This however may not be affordable since in practice the number of flows F is typically $\mathscr{O}(N^2)$, which can be too large. To cope with this curse of dimensionality, instead of $\{g_L^{(k)},g_Q^{(k)},g_B^{(k)},g_C^{(k)}\}$ judicious surrogates $\{\tilde{g}_L^{(k)},\tilde{g}_Q^{(k)},\tilde{g}_B^{(k)},\tilde{g}_C^{(k)}\}$, chosen based on the second-order Taylor-expansion around the previous updates, are minimized. As will be clear later, adopting these surrogates avoids inversion, and parallelizes the computations. The aforementioned surrogate for $g_L^{(k)}$ around $\mathbf{L}[k-1]$ is given as

$$\tilde{g}_L^{(k)}(\mathbf{L}):=g_L^{(k)}(\mathbf{L}[k-1])+\mathrm{tr}\big((\mathbf{L}-\mathbf{L}[k-1])'\nabla g_L^{(k)}(\mathbf{L}[k-1])\big) + \frac{\mu_L[k]}{2}\|\mathbf{L}-\mathbf{L}[k-1]\|_F^2 \quad (22)$$

for some $\mu_L[k]\geq\sigma_{\max}\big[\nabla^2 g_L^{(k)}(\mathbf{L}[k-1])\big]$ (likewise for $\tilde{g}_Q^{(k)},\tilde{g}_B^{(k)}$, and $\tilde{g}_C^{(k)}$). It is useful to recognize that each surrogate, say $\tilde{g}_L^{(k)}$, has the following properties: (i) it majorizes $g_L^{(k)}$, namely $g_L^{(k)}(\mathbf{L})\leq\tilde{g}_L^{(k)}(\mathbf{L}),\ \forall\mathbf{L}$; and it is locally tight, which means that (ii) $g_L^{(k)}(\mathbf{L}[k-1])=\tilde{g}_L^{(k)}(\mathbf{L}[k-1])$; and, (iii) $\nabla g_L^{(k)}(\mathbf{L}[k-1])=\nabla\tilde{g}_L^{(k)}(\mathbf{L}[k-1])$.

The sought approximation leads to an iterative procedure, where iteration k entails orderly updating $\{\mathbf{L}[k],\mathbf{Q}[k],\mathbf{B}[k],\mathbf{C}[k]\}$ by minimizing $\tilde{g}_L^{(k)},\tilde{g}_Q^{(k)},\tilde{g}_B^{(k)},\tilde{g}_C^{(k)}$, respectively; e.g., the update for $\mathbf{L}[k]$ is

$$\begin{aligned}\mathbf{L}[k]&=\arg\min_{\mathbf{L}\in\mathbb{R}^{F\times\rho}}\tilde{g}_L^{(k)}(\mathbf{L})\\&=\mathbf{L}[k-1]-(\mu_L[k])^{-1}\nabla g_L^{(k)}(\mathbf{L}[k-1])\end{aligned}$$

which is a nothing but a single step of gradient descent on $g_L^{(k)}$. Upon defining the residual matrices $\mathbf{\Phi}_y(\mathbf{L},\mathbf{Q},\mathbf{B},\mathbf{C}):=\mathbf{R}(\mathbf{L}\mathbf{Q}'+\mathbf{B}\odot\mathbf{C})-\mathbf{Y}$ and $\mathbf{\Phi}_z(\mathbf{L},\mathbf{Q},\mathbf{B},\mathbf{C}):=\mathscr{P}_\Pi(\mathbf{L}\mathbf{Q}'+\mathbf{B}\odot\mathbf{C})-\mathbf{Z}_\Pi$, the overall algorithm is listed in Table 2.

All in all, Algorithm 2 amounts to an iterative block-coordinate-descent scheme with four block updates per iteration, each minimizing a tight surrogate of (P5). Since each subproblem is smooth and strongly convex, the convergence follows from [36] as stated next.

Proposition 3: *[36] Upon choosing* $\{c_L^{'}\geq\mu_L[k]\geq\sigma_{\max}\big[\nabla^2 g_L^{(k)}(\mathbf{L}[k-1])\big]\}_{k=1}^{\infty}$ *for some* $c_L^{'}>0$ *(likewise for* $\mu_Q[k],\mu_B[k],\mu_C[k]$*), the iterates* $\{\mathbf{L}[k],\mathbf{Q}[k],\mathbf{B}[k],\mathbf{C}[k]\}$ *generated by Algorithm 2 converge to a stationary point of (P5).*

Remark 3 (Fast algorithms): In order to speed up the gradient descent iterations per block of Algorithm 2, Nesterov-type acceleration techniques along the lines of those introduced in e.g., [37] can be deployed, which can improve the $\mathscr{O}(1/k)$ convergence rate of the standard gradient descent to $\mathscr{O}(1/k^2)$.

Algorithm 2 : Alternating majorization-minimization solver for (P5)

input $\mathbf{Y},\mathbf{Z}_\Pi,\Pi,\mathbf{R},\mathbf{R}_L,\mathbf{R}_Q,\mathbf{R}_B,\mathbf{R}_C,\lambda_*,\lambda_1$, and $\{\mu_L[k],\mu_Q[k],\mu_B[k],\mu_C[k]\}_{k=1}^{\infty}$.
initialize $\mathbf{L}[0],\mathbf{Q}[0],\mathbf{B}[0],\mathbf{C}[0]$ at random, and set $k=0$.
while not converged **do**
[S1] Update L
$\mathbf{F}[k]=\mathbf{R}'\mathbf{\Phi}_y(\mathbf{L}[k],\mathbf{Q}[k],\mathbf{B}[k],\mathbf{C}[k])+\mathbf{\Phi}_z(\mathbf{L}[k],\mathbf{Q}[k],\mathbf{B}[k],\mathbf{C}[k])$
$\mathbf{L}[k+1]=\mathbf{L}[k]-\frac{1}{\mu_L[k]}\big(\mathbf{F}[k]\mathbf{Q}[k]+\lambda_*\mathbf{R}_L^{-1}\mathbf{L}[k]\big)$
[S2] Update Q
$\mathbf{G}[k]=\mathbf{\Phi}_y'(\mathbf{L}[k+1],\mathbf{Q}[k],\mathbf{B}[k],\mathbf{C}[k])\mathbf{R}+\mathbf{\Phi}_z'(\mathbf{L}[k+1],\mathbf{Q}[k],\mathbf{B}[k],\mathbf{C}[k])$
$\mathbf{Q}[k+1]=\mathbf{Q}[k]-\frac{1}{\mu_Q[k]}\Big[\mathbf{G}[k]\mathbf{L}[k+1]+\lambda_*\mathbf{R}_Q^{-1}\mathbf{Q}[k]\Big]$
[S3] Update B
$\mathbf{H}[k]=\mathbf{R}'\mathbf{\Phi}_y(\mathbf{L}[k+1],\mathbf{Q}[k+1],\mathbf{B}[k],\mathbf{C}[k])+\mathbf{\Phi}_z(\mathbf{L}[k+1],\mathbf{Q}[k+1],\mathbf{B}[k],\mathbf{C}[k])$
$\mathbf{B}[k+1]=\mathbf{B}[k]-\frac{1}{\mu_B[k]}\Big[\mathbf{C}[k]\odot\mathbf{H}[k]+\lambda_1\mathrm{unvec}\big(\mathbf{R}_{\mathrm{B}}^{-1}\mathrm{vec}(\mathbf{B}[\mathrm{k}])\big)\Big]$
[S4] Update C
$\mathbf{E}[k]=\mathbf{R}'\mathbf{\Phi}_y(\mathbf{L}[k+1],\mathbf{Q}[k+1],\mathbf{B}[k+1],\mathbf{C}[k])+\mathbf{\Phi}_y(\mathbf{L}[k+1],\mathbf{Q}[k+1],\mathbf{B}[k+1],\mathbf{C}[k])$
$\mathbf{C}[k+1]=\mathbf{C}[k]-\frac{1}{\mu_C[k]}\Big[\mathbf{B}[k]\odot\mathbf{E}[k]+\lambda_1\mathrm{unvec}\big(\mathbf{R}_{\mathrm{C}}^{-1}\mathrm{vec}(\mathbf{C}[\mathrm{k}])\big)\Big]$
$k\leftarrow k+1$
end while
return $(\mathbf{A}[k]=\mathbf{B}[k]\odot\mathbf{C}[k],\mathbf{X}[k]=\mathbf{L}[k]\mathbf{Q}'[k])$

VIII. Practical Considerations

Before assessing their relevance to large-scale networks, the proposed algorithms must address additional practical issues. Those relate to the fact that network data are typically decentralized, streaming, subject to outliers as well as misses, and the routing matrix may be either unknown or dynamically changing over time. This section sheds light on solutions to cope with such practical challenges.

A. Inconsistent partial measurements

Certain network links may not be easily accessible to collect measurements, or, their measurements might be lost during the communication process due to e.g., packet drops. Let Π_y collect the available link measurements during the time horizon $\mathscr{T}$. In addition, certain link or flow counts may not be consistent with the adopted model in (2) and (4). To account for possible presence of outliers introduce the matrices $\mathbf{O}_y \in \mathbb{R}^{L\times T}$ and $\mathbf{O}_z \in \mathbb{R}^{F\times T}$, which are nonzero at the positions associated with the outlying measurements, and zero elsewhere. The link-count model (2) should then be modified to $\mathbf{Y}_{\Pi_y} = \mathscr{P}_{\Pi_y}(\mathbf{R}(\mathbf{X}+\mathbf{A})+\mathbf{O}_y+\mathbf{V})$, and the flow counts to $\mathbf{Z}_\Pi = \mathscr{P}_\Pi(\mathbf{X}+\mathbf{A}+\mathbf{O}_z+\mathbf{W})$. Typically the outliers constitute a small fraction of measurements, thus rendering $\{\mathbf{O}_y, \mathbf{O}_z\}$ sparse. The optimization task (P1) can then be modified to take into account the misses and outliers as follows

$$\begin{aligned}\text{(P6)}\ (\hat{\mathbf{X}},\hat{\mathbf{A}}) = \arg\min_{\{\mathbf{X},\mathbf{A},\mathbf{O}_y,\mathbf{O}_z\}} & \frac{1}{2}\|\mathscr{P}_{\Pi_y}(\mathbf{Y}-\mathbf{R}(\mathbf{X}+\mathbf{A})-\mathbf{O}_y)\|_F^2 \\ & + \frac{1}{2}\|\mathscr{P}_\Pi(\mathbf{Z}-\mathbf{X}-\mathbf{A}-\mathbf{O}_z)\|_F^2 + \lambda_*\|\mathbf{X}\|_* + \lambda_1\|\mathbf{A}\|_1 \\ & + \lambda_y\|\mathbf{O}_y\|_1 + \lambda_z\|\mathbf{O}_z\|_1\end{aligned}$$

where λ_y and λ_z control the density of link- and flow-level outliers, respectively. Again, one can employ ADMM-type algorithms to solve (P6).

Routing information may not also be revealed in certain applications due to e.g., privacy reasons. In this case, each network link can potentially carry an unknown fraction of every OD flow. Let $\mathscr{L}_{\text{in}}(n)$ and $\mathscr{L}_{\text{out}}(n)$ denote the set of incoming and outgoing links to node $n \in \mathscr{N}$. The routing variables then must respect the flow conservation constraints, that is formally $\mathbf{R} \in \mathscr{R} := \{\mathbf{R} \in [0,1]^{L\times F} \ : \ \sum_{\ell\in\mathscr{L}_{\text{in}}(n)} r_{\ell,f} = \sum_{\ell\in\mathscr{L}_{\text{out}}(n)} r_{\ell,f},\ \forall f \in \mathscr{F}, n \in \mathscr{N}\}$. Taking the unknown routing variables into account, the optimization task to estimate the traffic is formulated as

$$\begin{aligned}\text{(P7)}\ \ (\hat{\mathbf{X}},\hat{\mathbf{A}}) = \arg\min_{\{\mathbf{X},\mathbf{A},\mathbf{R}\in\mathscr{R}\}} & \frac{1}{2}\|\mathbf{Y}-\mathbf{R}(\mathbf{X}+\mathbf{A})\|_F^2 \\ & + \frac{1}{2}\|\mathscr{P}_\Pi(\mathbf{Z}-\mathbf{X}-\mathbf{A})\|_F^2 + \lambda_*\|\mathbf{X}\|_* + \lambda_1\|\mathbf{A}\|_1\end{aligned}$$

which is nonconvex due to the presence of bilinear terms in the LS cost.

B. Real-time operation

Monitoring of large-scale IP networks necessitates collecting massive amounts of data which far outweigh the ability of modern computers to store and analyze them in real time. In addition, nonstationarities due to routing changes and missing data further challenges estimating traffic and anomalies. In dynamic networks routing tables are constantly readjusted to effect traffic load balancing and avoid congestion caused by e.g., traffic congestion anomalies or network infrastructure failures. On top of the previous arguments, in practice the measurements are acquired sequentially across time, which motivates updating previously obtained estimates rather than recomputing new ones from scratch each time a new datum becomes available.

To account for routing changes, let $\mathbf{R}_t \in \mathbb{R}^{L\times F}$ denote the routing matrix at time t. The observed link counts at time instant t then adhere to $\mathbf{y}_t = \mathbf{R}_t(\mathbf{x}_t + \mathbf{a}_t) + \mathbf{v}_t,\ t = 1, 2, \ldots$, where $\mathbf{y}_t \in \mathbb{R}^L$, and the partial flow counts at time t obey $\mathbf{z}_{\Pi_t} = \mathscr{P}_{\Pi_t}(\mathbf{x}_t + \mathbf{a}_t + \mathbf{w}_t),\ t = 1, 2, \ldots$, where $\mathbf{z}_{\Pi_t} \in \mathbb{R}^F$, and Π_t indexes the OD flows measured at time t. In order to estimate the nominal and anomalous traffic components $(\mathbf{x}_t, \mathbf{a}_t)$ at time instant t in real time, given only the past observations $\{\mathbf{y}_\tau, \mathbf{z}_{\Pi_\tau}\}_{\tau=1}^t$, the framework developed in our companion paper [16] can be adopted. Building on the fact that the traffic traces $\{\mathbf{x}_t\}_{t=1}^\infty$ lie in a low-dimensional linear subspace, say $\mathscr{L}$, one can postulate $\mathbf{x}_t = \mathbf{L}\mathbf{q}_t$ for $\mathbf{L} \in \mathbb{R}^{F\times\rho}$ with $\rho \ll F$, where $\mathbf{L}$ spans the subspace $\mathscr{L}$. Pursuing the ideas in [16], the nuclear-norm characterization in (18), which enjoys separability across time, can be applied to formulate exponentially-weighted LS estimators. The corresponding optimization task can then be solved via alternating minimization algorithms [16].

It is worth commenting that the companion work [16] aims primarily at identifying the anomalies $\mathbf{a}_t$ from link counts, which requires slow variations of the routing matrix to ensure $\{\mathbf{R}_t\mathbf{x}_t\}_{t=1}^\infty$ lie in a low-dimensional subspace. However, the tomography task considered in the present paper imposes no restriction on the routing matrix. Indeed, routing variability helps estimation of the nominal traffic $\mathbf{x}_t$. More precisely, suppose that $\{\mathbf{R}_t\}$ are sufficiently distinct so as the intersection of the nullspaces $\bigcap_t \mathscr{N}_{R_t}$ has a small dimension. Consequently, it is less likely to find an alternative feasible solution $\mathbf{X}_1 := \mathbf{X}_0 + \mathbf{H}$ with $\mathbf{H} := [\mathbf{h}_1, \ldots, \mathbf{h}_F]$ and $\mathbf{h}_t \in \mathscr{N}_{R_t}$ such that $\mathbf{H} \in \Phi_{X_0}$ (cf. Section IV); see also Lemma 1. Further analysis of this intriguing phenomenon goes beyond the scope of the present paper, and will be pursued as future research.

C. Decentralized implementation

Algorithms 1 and 2 demand each network node (router) $n \in \mathscr{N}$ continuously communicate the local measurements of its incident links as well as the OD-flow counts originating at node n, to a central monitoring station. While this is typically the prevailing operational paradigm adopted in current network technologies, there are limitations associated with this architecture. Collecting all these data at the routers may lead to excessive protocol overhead, especially for large-scale networks with high acquisition rate. In addition, with the exchange of raw measurements missing data due to communication errors are inevitable. Performing the optimization in a centralized fashion raises robustness concerns as well, since the central monitoring station represents an isolated point of failure.

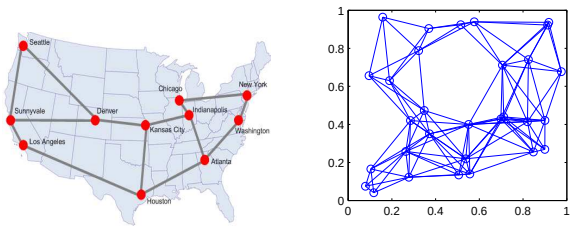
The aforementioned reasons motivate devising fully distributed iterative algorithms in large-scale networks, which allocate the network tomography functionality to the routers. In a nutshell, per iteration, nodes carry out simple computational tasks locally, relying on their own local measurements. Subsequently, local estimates are refined after exchanging messages only with directly connected neighbors, which facilitates percolation of information to the entire network. The ultimate goal is for the network nodes to consent on the global map of network-traffic-state $(\hat{\mathbf{X}}, \hat{\mathbf{A}})$, which remains close to the one obtained via the centralized counterpart with the entire network data available at once. Building on the separable characterization of the nuclear norm in (18), and adopting ADMM method as a basic tool to carry out distributed optimization, a generic framework for decentralized sparsity-regularized rank minimization was put forth in our companion paper [15]. In the context of network anomaly detection, the results there are encouraging and the proposed ideas can be applied to solve also (P1) in a distributed fashion.

IX. Performance Evaluation

Performance of the novel schemes is assessed in this section via computer simulations with both synthetic and real network data as described below.

Synthetic network data. The network topology is generated according to a random geometric graph model, where the nodes are randomly placed in a unit square, and two nodes are connected with an edge if their distance is less than a prescribed threshold d_c. In general, to form the routing matrix each OD pair takes K nonoverlapping paths, each determined according to the minimum hop-count algorithm. After finding the routes, links carrying no traffic are discarded. Clearly, the number of links varies according to d_c. The underlying traffic matrix $\mathbf{X}_0$ follows the bilinear model $\mathbf{X}_0 = \mathbf{L}\mathbf{Q}'$, with the factors $\mathbf{L} \in \mathbb{R}^{F\times\rho}$ and $\mathbf{Q} \in \mathbb{R}^{T\times\rho}$ having i.i.d. Gaussian entries $\mathcal{N}(0, 1/F)$ and $\mathcal{N}(0, 1/T)$, respectively. Entries of the anomaly matrix $\mathbf{A}_0$ are also randomly drawn from the set $\{-1, 0, 1\}$ with probability (w.p.) $\Pr(a_{f,t} = -1) = \Pr(a_{f,t} = 1) = p/2$, and $\Pr(a_{f,t} = 0) = 1 - p$. The link loads are then formed as $\mathbf{Y} = \mathbf{R}(\mathbf{X}_0 + \mathbf{A}_0)$. A subset of OD flows is also sampled uniformly at random to form the partial OD flow-level measurements $\mathbf{Z}_\Pi = \mathbf{\Pi} \odot (\mathbf{X}_0 + \mathbf{A}_0)$, where each entry of $\mathbf{\Pi} \in \{0, 1\}^{F\times T}$ is i.i.d. Bernoulli distributed taking value one w.p. π, and zero w.p. $1 - \pi$.

Real network data. Real data including OD flow traffic levels are collected from the operation of the Internet-2 network (Internet backbone network across USA) [17], shown in Fig. 3 (a). Internet-2 comprises $N = 11$ nodes, $L = 41$ links, and $F = 121$ OD flows. Flow traffic levels are recorded every five-minute interval, for a three-week operational period during December 8-28, 2003 [17], [38]. The collected flow levels are the aggregation of clean and anomalous traffic components, that is sum of unknown "ground-truth" low-rank and sparse matrices $\mathbf{X}_0 + \mathbf{A}_0$. The "ground truth" components are then discerned from their aggregate after applying robust PCP algorithms developed e.g., in [25]. The recovered $\mathbf{X}_0$ exhibits three dominant singular values, confirming the low-rank property of the nominal traffic matrix. Also, after retaining only the significant spikes with magnitude larger than the threshold $50\|\mathbf{Y}\|_F/LT$, the formed anomaly matrix $\mathbf{A}_0$ has 1.10% nonzero entries. The link loads in $\mathbf{Y}$ are obtained through multiplication of the aggregate traffic with the Internet-2 routing matrix. Even though $\mathbf{Y}$ is "constructed" here from flow measurements, link loads are acquired from SNMP traces [39]. Moreover, the aggregate flow traffic matrix $\mathbf{X}_0+\mathbf{A}_0$ is sampled uniformly at random with probability π to form $\mathbf{Z}_\Pi$. In practice, these samples are acquired via NetFlow protocol [7].



(a) (b)

Fig. 3. Network topology graphs. (a) Internet-2. (b) Random synthetic network with $N = 30$ and $d_c = 0.35$.

A. Exact recovery validation

To demonstrate the merits of (P2) in accurately recovering the true values $(\mathbf{X}_0, \mathbf{A}_0)$, it is solved for a wide range of rank r and (average) sparsity levels $s = pFT$ using the ADMM solver in Algorithm 1. Synthetic data is generated as described before for a random network with $N = 30$, $d_c = 0.35$, and $F = T = N(N - 1)/3$; see Fig. 3(b). For F randomly selected OD pairs, K nonoverlapping paths are chosen to carry the traffic. Each path is created based on the minimum-hop count routing algorithm to form the routing matrix. A random fraction of the origin's traffic is also assigned to each path. The gray-scale plots in Fig. 4 show phase transition for the relative estimation error $e_{x+a} = e_x + e_a$, including both nominal $e_x := \|\hat{\mathbf{X}} - \mathbf{X}_0\|_F/\|\mathbf{X}_0\|_F$, and anomalous traffic estimation error $e_a := \|\hat{\mathbf{A}} - \mathbf{A}_0\|_F/\|\mathbf{A}_0\|_F$ under various percentage of misses. Top figure is associated with $K = 1$, while for the bottom figure $K = 3$. The parameter λ in (P2) is also tuned to optimize the performance.

When single-path routing is used, the network entails $L = 159$ physical links. In this case, the routing matrix $\mathbf{R} \in \{0, 1\}^{159\times 290}$ has a huge nullspace with $\dim(\mathcal{N}_R) = 127$, and as a result Fig. 4 (top) indicates that accurate recovery is possible only for relatively small values of r and s. However, when multipath routing ($K = 3$) is used, there are more $L = 227$ physical links involved in carrying the traffic of OD flows. This shrinks the nullspace of $\mathbf{R} \in [0, 1]^{227\times 290}$ to $\dim(\mathbf{R}) = 68$, and improves the isometry property of $\mathbf{R}$ for sparse vectors. As a result, under traffic of higher dimensionality and denser anomalies accurate traffic estimation is possible; see Fig. 4 (bottom).

B. Traffic and anomaly maps

Real Internet-2 data is considered to portray the traffic based on (P1) every 42-hour interval, which amounts to time horizon

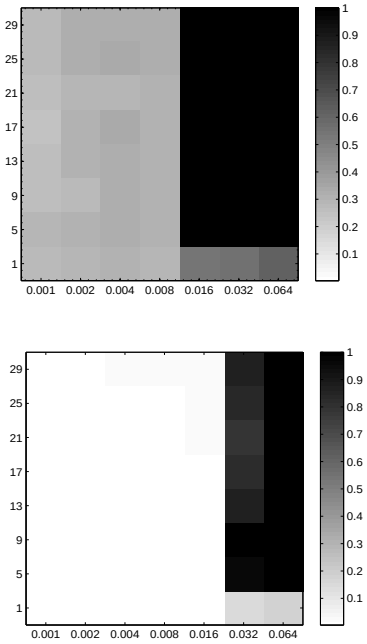


Fig. 4. Relative estimation error e_{x+a} for various values of rank (r) and sparsity level ($s = pFT$) where $F = T = 290$ and $\pi = 0.25$. (a) Single-path routing versus (b) multipath routing ($K = 3$). White represents exact recovery ($e_{x+a} \approx 0$), while black represents $e_{x+a} \approx 1$.

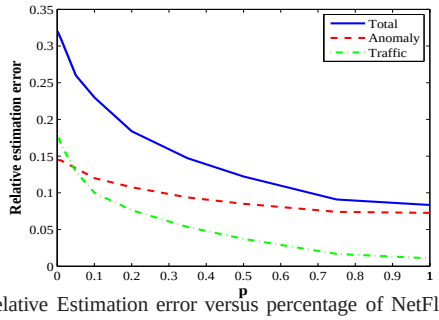
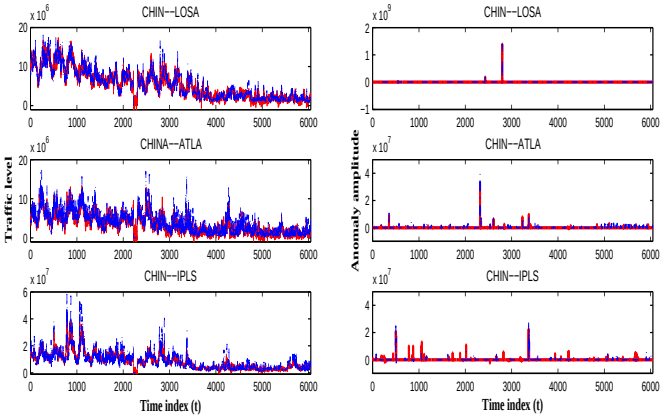


Fig. 5. Relative Estimation error versus percentage of NetFlow samples.



(a) (b)

Fig. 6. Nominal (a) and anomalous (b) traffic portrays for three representative OD flows when $\pi = 0.1$. True traffic is dashed blue and the estimated one is solid red.

of $T = 504$ time bins.

Impact of NetFlow data. The role of NetFlow measurements on the traffic estimation performance is depicted in Fig. 5 plotting the relative error e_{x+a} for various percentages of NetFlow samples (π). Normally, the estimation accuracy improves as π grows, where the improvement seems more pronounced for the nominal traffic. When only the link loads are available, adding 10% NetFlow samples enhances the nominal-traffic estimation accuracy by 45%, while the one for the anomalous traffic is improved by 18%. This observation corroborates the effectiveness of exploiting partial NetFlow samples toward mapping out the network traffic.

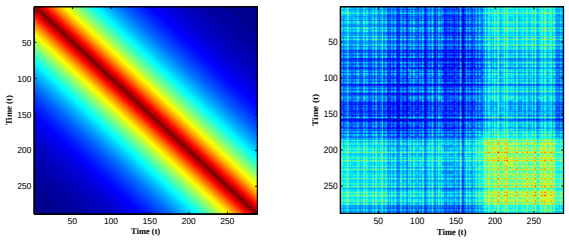
Traffic profiles. For $\pi = 0.1$, the true and estimated traffic time-series are illustrated in Fig. 6 for three representative OD flows originating from the CHIN autonomous system located at Chicago. The depicted time-series correspond to three different rows of $\hat{\mathbf{X}}$ and $\hat{\mathbf{A}}$ returned by (P1). It is apparent that the traffic variations are closely tracked and significant spikes are correctly picked by (P1). It pinpoints confidently a significant anomaly occurring within 9:20 P.M.–9:25 P.M., December 11, 2003, in the flow CHIN–LOSA, which traverses several physical links. High false alarm declared for the CHIN–IPLS flow is also because it visits only a single link, and thus not revealing enough information.

Unveiling anomalies. Identifying anomalous patterns is pivotal towards proactive network security tasks. The resultant estimated map $\hat{\mathbf{A}}$ returned by (P1) offers a depiction of the network health-state across both time and flows. Our previous work in [12] and [16] deals with creating such a map with only the link loads $\mathbf{Y}$ at hand (i.e., $\Pi = \emptyset$), and the primary goal is to recover $\hat{\mathbf{A}}$. The purported results in [12], [16] are promising and could markedly outperform state-of-art workhorse PCA-based approaches in e.g., [13], [40]. Relative to [12], [16], the current work however allows additional partial flow-level measurements. This naturally raises the question how effective this additional information is toward identifying the anomalies. As seen in Fig. 5, taking more NetFlow samples is useful, but beyond a certain threshold it does not offer any extra appeal.

C. Estimation with spatiotemporal correlation information

This section evaluates the effectiveness of (P5) and demonstrates the usefulness of traffic correlation information. Training data from the week December 8-15, 2003 are used to estimate the Internet-2 traffic on the next day, December 16, 2003. The nominal "ground truth" traffic matrix $\mathbf{X}_0$ described earlier is considered, and for validation purposes bursty anomalies are synthetically injected to form the aggregate traffic $\mathbf{X}_0 + \mathbf{A}_0$, which is then used to generate $\mathbf{Y}$ and $\mathbf{Z}_\Pi$. To simulate the NetFlow samples, suppose 10% of randomly selected OD flows are inaccessible for the entire time horizon, and the rest are sampled only 10% of time, resulting in 9% flow-level measurements available.

Bursty anomalies. To generate anomalies $\mathbf{X}_0$, envision a scenario where a subset of OD flows undergo bursty anomalies while the rest are clean. Per flow f bursty anomalies are generated according to the random multiplicative process $\{a_{f,t} = \gamma_f b_{f,t} c_{f,t}\}_t$, with mutually independent stationary processes $\{c_{f,t}\}$ and $\{b_{f,t}\}$. The former is a correlated Gaussian process, and the latter is a correlated $\{0,1\}$-Bernoulli



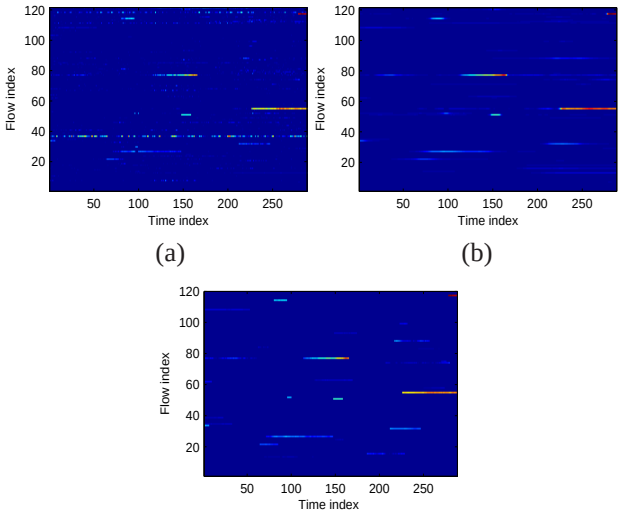
(a) (b)

Fig. 7. Sample correlations $\mathbf{R}_B$ (a) and $\mathbf{R}_Q$ (b) learned based on historical traffic data during December 8-15, 2003.

process to model the bursts. The Gaussian process obeys the first-order auto-regressive model $c_{f,t} = \theta c_{f,t-1} + \sigma_n n_{f,t}$, with $c_{f,0} = 0$ and $n_{f,t} \sim \mathscr{N}(0,1)$ for some $\theta < 1$. The Bernoulli process also adheres to $b_{f,t} = d_{f,t} b_{f,t-1} + (1 - d_{f,t}) e_{f,t}$, where the independent random variables $d_{f,t}$ and $e_{f,t}$ obey $d_{f,t} \sim \text{Ber}(\alpha)$ and $e_{f,t} \sim \text{Ber}(\nu)$, respectively. Initial variable $b_{f,0}$ is also generated as $\text{Ber}(\nu)$.

Learning correlations. Owing to the stationarity of processes $\{b_{f,t}\}$ and $\{c_{f,t}\}$, process $\{a_{f,t}\}$ is stationary, and as a result $R_a^{(f)}(\tau) = \gamma_f^2 R_b^{(f)}(\tau) R_c^{(f)}(\tau)$, with the corresponding correlations given as $R_c^{(f)}(\tau) = \theta^\tau \sigma_n^2/(1-\theta^2)$ and $R_b^{(f)}(\tau) = \nu(1-\nu)\alpha^\tau + \nu$. Set $\gamma_f = 50$, $\theta = 0.999$, $\sigma_n = 0.005$, $\alpha = 0.98$, and $\nu = 0.03$. The correlation matrices $\{\mathbf{R}_B, \mathbf{R}_C\}$ with Toeplitz blocks are then obtained from (20). Moreover, to account for the cyclostationarity of traffic with a day-long periodicity, the correlation matrices $\{\mathbf{R}_L, \mathbf{R}_Q\}$ are learned as elaborated in Section VI-A. The resulting temporal correlation matrices $\mathbf{R}_B$ and $\mathbf{R}_Q$, learned based on the traffic data December 8-15, 2003, are displayed in Fig. 7, where 288 data points in each axis correspond to 24 hours. The sharp transition noticed in Fig. 7 (b) happens at $3:45$ p.m. that signifies a sudden increase in the traffic usage for the rest of the day.

Traffic maps. Fig. 9 depicts the time series of estimated and true nominal traffic for the IPLS–CHIN OD flow (see Fig. 3(a)). For this flow, no direct NetFlow sample is collected. It is apparent that (P5) which uses the knowledge of traffic spatiotemporal correlation tracks fairly well the underlying traffic, whereas (P1) cannot even track the large-scale variations of traffic. This demonstrates the nonidentifiability of (P1) when only a small fraction 9% of OD flows are nonuniformly sampled, and notably around 10% of rows of $\mathbf{X}_0$ are not directly observable. (P5) however interpolates the traffic associated with unobserved OD flows with the observed ones through the correlation matrices $\{\mathbf{R}_L, \mathbf{R}_Q\}$. The resulting relative estimation error for (P5) is $e_x = 0.19$, which is well below $e_x = 0.62$ for (P1). The correlation knowledge also helps discovering the anomalous traffic patterns as seen from Fig. 8, where in particular (P5) attains $e_a = 0.27$, while (P1) does $e_a = 0.73$. Interestingly, the anomaly map revealed by (P1) tends to spot the anomalies intermittently since the ℓ_1-norm regularizer weighs all flows and time-instants equally.



(a) (b)

(c)

Fig. 8. Estimated and "ground truth" (c) anomaly maps across time and flows without using correlation (a), and after using correlation information (b).

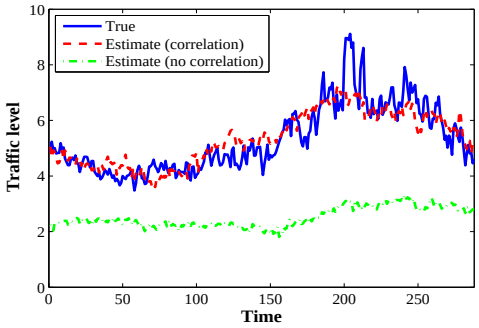


Fig. 9. True and estimated traffic of IPLS-CHIN flow.

X. Conclusions and Future Work

This paper taps on recent advances in low-rank and sparse recovery to create maps of nonmimal and anomalous traffic as a valuable input for network management and proactive security tasks. A novel tomographic framework is put forth which subsumes critical network monitoring tasks including traffic estimation, anomaly identification, and traffic interpolation. Leveraging low intrinsic-dimensionality of nominal traffic as well as the sparsity of anomalies, a convex program is formulated with ℓ_1- and nuclear-norm regularizers, with the link loads and a small subsets of flow counts as the available data. Under certain circumstances on the true traffic and anomalies in addition to the routing and OD-flow sampling strategies, sufficient conditions are derived, which guarantee accurate estimation of the traffic.

For practical networks where the said conditions are possibly violated, additional knowledge about inherent traffic patterns are incorporated through correlations by adopting a Bayesian approach and taking advantage of the bilinear characterization of the ℓ_1- and nuclear-norm. A systematic approach is also devised to learn the correlations using (cyclo)stationary historical traffic data. Simulated tests with synthetic and real

Internet data confirm the efficacy of the novel estimators. There are yet intriguing unanswered questions that go beyond the scope of the current paper, but worth pursuing as future research. One such question pertains to quantifying a minimal count of sampled OD flows for a realistic network scenario with a given routing matrix, which assures accurate traffic estimation. Another avenue to explore involves adoption of tensor models along the lines of [33], [41], [42] to further exploit the network topological information toward improving the traffic estimation accuracy.

APPENDIX
PROOF OF THE MAIN RESULT

In what follows, conditions are first derived under which the pair $(\mathbf{X}_0, \mathbf{A}_0)$ is the *unique* optimal solution of (P2). The sought conditions pertain to existence of certain dual certificates, which are then constructed in Section B.

A. Unique Optimality Conditions

Recall the *nonsmooth* optimization problem (P2), and its Lagrangian formed as

$$\begin{aligned}\mathscr{L}(\mathbf{X},\mathbf{A};\mathbf{M}_y,\mathbf{M}_z) =&\|\mathbf{X}\|_*+\lambda\|\mathbf{A}\|_1+\langle\mathbf{M}_y,\mathbf{Y}-\mathbf{R}(\mathbf{X}+\mathbf{A})\rangle\\&+\langle\mathbf{M}_z,\mathbf{Z}_\Pi-\mathscr{P}_\Pi(\mathbf{X}+\mathbf{A})\rangle \qquad (23)\end{aligned}$$

where $\mathbf{M}_y\in\mathbb{R}^{L\times T}$ and $\mathbf{M}_z\in\mathbb{R}^{F\times T}$ are the matrices of dual variables (multipliers) associated with the link and flow level constraints in (P2), respectively. From the characterization of the subdifferential for the nuclear- and the ℓ_1-norm (see e.g., [43]), the subdifferential of the Lagrangian at $(\mathbf{X}_0,\mathbf{A}_0)$ is given by (recall that $\mathbf{X}_0=\mathbf{U}_0\mathbf{\Sigma}_0\mathbf{V}_0'$)

$$\begin{aligned}\partial_{\mathbf{X}}\mathscr{L}(\mathbf{X}_0,\mathbf{A}_0;\mathbf{M}_y,\mathbf{M}_z)=\Big\{&\mathbf{U}_0\mathbf{V}_0^{'}+\mathbf{W}-\mathbf{R}'\mathbf{M}_y-\mathscr{P}_\Pi(\mathbf{M}_z):\\&\|\mathbf{W}\|\le 1,\ \mathscr{P}_{\Phi_{X_0}}(\mathbf{W})=\mathbf{0}_{F\times T}\Big\} \qquad (24)\end{aligned}$$

$$\begin{aligned}\partial_{\mathbf{A}}\mathscr{L}(\mathbf{X}_0,\mathbf{A}_0;\mathbf{M}_y,\mathbf{M}_z)=\{&\lambda\mathrm{sign}(\mathbf{A}_0)+\lambda\mathbf{F}-\mathbf{R}'\mathbf{M}_y-\mathscr{P}_\Pi(\mathbf{M}_z):\\&\|\mathbf{F}\|_\infty\le 1,\ \mathscr{P}_{\Omega_{A_0}}(\mathbf{F})=\mathbf{0}_{F\times T}\}. \qquad (25)\end{aligned}$$

The optimality conditions for (P2) assert that $(\mathbf{X}_0,\mathbf{A}_0)$ is an optimal (not necessarily unique) solution if and only if

$$\begin{aligned}\mathbf{0}_{F\times T}&\in\partial_{\mathbf{A}}\mathscr{L}(\mathbf{X}_0,\mathbf{A}_0;\mathbf{M}_y,\mathbf{M}_z)\\ \mathbf{0}_{F\times T}&\in\partial_{\mathbf{X}}\mathscr{L}(\mathbf{X}_0,\mathbf{A}_0;\mathbf{M}_y,\mathbf{M}_z).\end{aligned}$$

This is tantamount to existence of the dual variables $\{\mathbf{W},\mathbf{F},\mathbf{M}_y,\mathbf{M}_z\}$ satisfying: (i) $\|\mathbf{W}\|\le 1$, $\mathscr{P}_{\Phi_{X_0}}(\mathbf{W})=\mathbf{0}_{F\times T}$, (ii) $\|\mathbf{F}\|_\infty\le 1$, $\mathscr{P}_{\Omega_{A_0}}(\mathbf{F})=\mathbf{0}_{F\times T}$, and (iii) $\lambda\mathrm{sign}(\mathbf{A}_0)+\lambda\mathbf{F}=\mathbf{U}\mathbf{V}'+\mathbf{W}=\mathbf{R}'\mathbf{M}_y-\mathscr{P}_\Pi(\mathbf{M}_z)$.

In essence, to eliminate $\mathbf{M}_y,\mathbf{M}_z$, one can alternatively interpret iii) as finding the dual variable $\mathbf{\Gamma}\in\mathscr{N}_R^\perp+\mathscr{N}_\Pi^\perp=(\mathscr{N}_R\cap\mathscr{N}_\Pi)^\perp$ such that $\mathbf{\Gamma}=\lambda\mathrm{sign}(\mathbf{A}_0)+\lambda\mathbf{F}=\mathbf{U}\mathbf{V}'+\mathbf{W}$. Since $\mathbf{W}=\mathscr{P}_{\Phi_{X_0}^\perp}(\mathbf{\Gamma})$ and $\mathbf{F}=\mathscr{P}_{\Omega_{A_0}^\perp}(\mathbf{\Gamma})$, conditions (i) and (ii) can also be simply recast in terms of $\mathbf{\Gamma}$. In general, (i)–(iii) may hold for multiple solution pairs. However, the next lemma asserts that a slight tightening of the optimality conditions (i)–(iii) leads to a *unique* optimal solution for (P2). The proof goes along the lines of [12, Lemma 2], and it is omitted here for conciseness.

Proposition 4: *If $(\mathbf{X}_0,\mathbf{A}_0)$ is locally identifiable from (c1) and (c2), and there exists a dual certificate $\mathbf{\Gamma}\in\mathbb{R}^{F\times T}$ satisfying*

$$\begin{aligned}&\mathrm{C1)}\quad \mathscr{P}_{\Phi_{X_0}}(\mathbf{\Gamma})=\mathbf{U}_0\mathbf{V}_0'\\&\mathrm{C2)}\quad \mathscr{P}_{\Omega_{A_0}}(\mathbf{\Gamma})=\lambda\mathrm{sign}(\mathbf{A}_0)\\&\mathrm{C3)}\quad \mathscr{P}_{\mathscr{N}_R\cap\mathscr{N}_\Pi}(\mathbf{\Gamma})=\mathbf{0}\\&\mathrm{C4)}\quad \|\mathscr{P}_{\Phi_{X_0}^\perp}(\mathbf{\Gamma})\|<1\\&\mathrm{C5)}\quad \|\mathscr{P}_{\Omega_{A_0}^\perp}(\mathbf{\Gamma})\|_\infty<\lambda\end{aligned}$$

then $(\mathbf{X}_0,\mathbf{A}_0)$ is the unique optimal solution to (P2).

The rest of the proof deals with construction of a valid dual certificate $\mathbf{\Gamma}$ that simultaneously meets C1–C5.

One should note that condition (iii) is a distinct feature of the recovery task pursued in this paper. In a similar context, in the robust PCP problem studied in [14], $\mathscr{N}_R=\emptyset,\mathscr{N}_\Pi=\emptyset$, and thus C3 does not appear anymore. In addition, the low-rank plus compressed sparse recovery task studied in [12] does not involve the intersection of subspaces as appearing in C3.

B. Dual Certificate Construction

The main steps of the construction are inspired by [14] which studies decomposition of low-rank plus sparse matrices, that is, $\Pi=\emptyset$ and $\mathbf{R}=\mathbf{I}_F$. However, relative to [14] the problem here brings up several new distinct elements including the null space of compression and sampling operators in C3, which further challenge construction of dual certificates, and demands, in part, a new treatment. In addition, different incoherence measures are introduced here which facilitate satisfiability for random ensembles. The construction involves two steps. In the first step, a candidate dual certificate is selected to fulfil C1–C3, whereas the second step assures the candidate dual certificate satisfies C4–C5 as well under certain technical conditions in terms of the incoherence parameters in Section IV-B.

Toward the first step, condition (II) in Theorem 1 implies local identifiability of the observation model, namely $\Omega_{A_0}\cap\Phi_{X_0}=\{\mathbf{0}\}$ and $(\Omega_{A_0}\oplus\Phi_{X_0})\cap(\mathscr{N}_R\cap\mathscr{N}_\Pi)=\{\mathbf{0}\}$, and thus based on a property of direct-sum [24] there *exists* a *unique* certificate $\mathbf{\Gamma}\in\Omega_{A_0}\oplus\Phi_{X_0}\oplus(\mathscr{N}_R\cap\mathscr{N}_\Omega)$ with projections $\mathscr{P}_{\Omega_{A_0}}(\mathbf{\Gamma})=\lambda\mathrm{sign}(\mathbf{A}_0)$, $\mathscr{P}_{\Phi_{X_0}}(\mathbf{\Gamma})=\mathbf{U}_0\mathbf{V}_0'$, and $\mathscr{P}_{\mathscr{N}_R\cap\mathscr{N}_\Pi}(\mathbf{\Gamma})=\mathbf{0}$. This dual certificate can be expressed as $\mathbf{\Gamma}=\mathbf{\Gamma}_{\Omega_{A_0}}+\mathbf{\Gamma}_{\Phi_{X_0}}+\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi}$ with the components $\mathbf{\Gamma}_{\Omega_{A_0}}\in\Omega_{A_0}$, $\mathbf{\Gamma}_{\Phi_{X_0}}\in\Phi_{X_0}$, and $\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi}\in\mathscr{N}_R\cap\mathscr{N}_\Pi$. As will be seen later, it is more convenient to represent $\mathbf{\Gamma}_{\Omega_{A_0}}=\epsilon_{\Omega_{A_0}}+\lambda\mathrm{sign}(\mathbf{A}_0)$ and $\mathbf{\Gamma}_{\Phi_{X_0}}=\epsilon_{\Phi_{X_0}}+\mathbf{U}_0\mathbf{V}_0'$. From C1–C3, for the projection components $\{\epsilon_{\Omega_{A_0}},\epsilon_{\Phi_{X_0}},\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi}\}$ it then holds that

$$\epsilon_{\Phi_{X_0}}=-\mathscr{P}_{\Phi_{X_0}}(\epsilon_{\Omega_{A_0}}+\lambda\mathrm{sign}(\mathbf{A}_0)+\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi}) \qquad (26)$$

$$\epsilon_{\Omega_{A_0}}=-\mathscr{P}_{\Omega_{A_0}}(\epsilon_{\Phi_{X_0}}+\mathbf{U}_0\mathbf{V}_0'+\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi}) \qquad (27)$$

$$\begin{aligned}&\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi}\\&=-\mathscr{P}_{\mathscr{N}_R\cap\mathscr{N}_\Pi}(\epsilon_{\Phi_{X_0}}+\mathbf{U}_0\mathbf{V}_0'+\epsilon_{\Omega_{A_0}}+\lambda\mathrm{sign}(\mathbf{A}_0)). \qquad (28)\end{aligned}$$

The second step of the proof manages the candidate dual certificate $\mathbf{\Gamma}$ to satisfy C4 and C5 as well. The main idea is to tighten the conditions for local identifiability, and impose

additional conditions on the incoherence measures (c.f. Section IV-B) to ensure that C4 and C5 hold true. In this direction, one can begin by bounding

$$\begin{aligned}\|\mathscr{P}_{\Phi^{\perp}_{X_0}}(\mathbf{\Gamma})\| &\leq \|\mathbf{\Gamma}_{\Omega_{A_0}} + \mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Omega}\| \\ &= \|\epsilon_{\Omega_{A_0}} + \lambda\mathrm{sgn}(\mathbf{A}_0) + \mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Omega}\| \\ &\overset{(a)}{\leq} \|\mathscr{P}_{\Omega_{A_0}}(\epsilon_{\Phi_{X_0}} + \mathbf{U}_0\mathbf{V}_0' + \mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Omega})\| \\ &\quad + \lambda\|\mathrm{sgn}(\mathbf{A}_0)\| + \|\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Omega}\| \end{aligned} \tag{29}$$

and

$$\begin{aligned}\|\mathscr{P}_{\Omega^{\perp}_{A_0}}(\mathbf{\Gamma})\|_\infty &\leq \|\mathbf{\Gamma}_{\Phi_{X_0}} + \mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi}\|_\infty \\ &= \|\epsilon_{\Phi_{X_0}} + \mathbf{U}_0\mathbf{V}_0^{'} + \mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi}\|_\infty \\ &\overset{(b)}{\leq} \|\mathscr{P}_{\Phi_{X_0}}(\epsilon_{\Omega_{A_0}} + \lambda\mathrm{sgn}(\mathbf{A}_0) + \mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi})\|_\infty \\ &\quad + \|\mathbf{U}_0\mathbf{V}_0^{'}\|_\infty + \|\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi}\|_\infty \end{aligned} \tag{30}$$

where (a) and (b) come from (26) and (27) after applying the triangle inequality. In order to bound the r.h.s. of (29) and (30), it is instructive first to recognize that $\|\mathbf{U}_0\mathbf{V}_0^{'}\|_\infty \leq \gamma(\mathbf{U}_0,\mathbf{V}_0)$, and

$$\|\mathrm{sgn}(\mathbf{A}_0)\| \leq \left(\|\mathrm{sgn}(\mathbf{A}_0)\|_{\infty,\infty}\|\mathrm{sgn}(\mathbf{A}_0)\|_{1,1}\right)^{1/2} = k \tag{31}$$

see e.g., [24]. In addition, building on (8) and (12), the first term in the r.h.s. of (29) is bounded as

$$\begin{aligned}&\|\mathscr{P}_{\Omega_{A_0}}(\epsilon_{\Phi_{X_0}} + \mathbf{U}_0\mathbf{V}_0^{'} + \mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi})\| \\ &\leq \|\mathscr{P}_{\Omega_{A_0}}\mathscr{P}_{\Phi_{X_0}}(\epsilon_{\Phi_{X_0}} + \mathbf{U}_0\mathbf{V}_0^{'})\| \\ &\qquad + \|\mathscr{P}_{\Omega_{A_0}}\mathscr{P}_{\mathscr{N}_\Pi}\mathscr{P}_{\mathscr{N}_R}(\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi})\| \\ &\overset{(a)}{\leq} \mu(\Phi_{X_0},\Omega_{A_0})\left(\|\epsilon_{\Phi_{X_0}}\| + 1\right) \\ &\qquad + \mu(\mathscr{N}_R,\Omega_{A_0}\cap\mathscr{N}_\Pi)\|\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Omega}\| \end{aligned} \tag{32}$$

where (a) is due to the fact that $\mathscr{P}_{\Omega_{A_0}\cap\mathscr{N}_\Pi} = \mathscr{P}_{\Omega_{A_0}}\mathscr{P}_{\mathscr{N}_\Pi}$.

Proceeding in a similar manner as for (32), upon using (11) it follows that

$$\begin{aligned}&\|\mathscr{P}_{\Phi_{X_0}}(\epsilon_{\Omega_{A_0}} + \lambda\mathrm{sgn}(\mathbf{A}_0) + \mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi})\|_\infty \\ &\leq \gamma(\mathbf{U}_0,\mathbf{V}_0)\ \|\mathscr{P}_{\Phi_{X_0}}(\epsilon_{\Omega_{A_0}} + \lambda\mathrm{sgn}(\mathbf{A}_0) + \mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi})\| \\ &\leq \gamma(\mathbf{U}_0,\mathbf{V}_0)\ \big[\|\mathscr{P}_{\Phi_{X_0}}(\epsilon_{\Omega_{A_0}} + \lambda\mathrm{sgn}(\mathbf{A}_0)\| \\ &\qquad + \|\mathscr{P}_{\Phi_{X_0}}\mathscr{P}_{\mathscr{N}_\Pi}(\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi})\|\big] \\ &\leq \gamma(\mathbf{U}_0,\mathbf{V}_0)\big[\mu(\Omega_{A_0},\Phi_{X_0})\left(\|\epsilon_{\Omega_{A_0}}\| + \lambda k\right) \\ &\qquad + \mu(\Phi_{X_0},\mathscr{N}_\Pi)\|\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi}\|\big]. \end{aligned} \tag{33}$$

Focusing on (33) and (32), it is only left to bound $\|\epsilon_{\Omega_{A_0}}\|$, $\|\epsilon_{\Phi_{X_0}}\|$, and $\|\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Omega}\|$. To this end, (27)-(28) are utilized to arrive at

$$\begin{aligned}&\|\epsilon_{\Phi_{X_0}}\| = \|\mathscr{P}_{\Phi_{X_0}}(\epsilon_{\Omega_{A_0}} + \lambda\mathrm{sign}(\mathbf{A}_0) + \mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi})\| \\ &\leq \mu(\Phi_{X_0},\Omega_{A_0})\left(\|\epsilon_{\Omega_{A_0}}\| + \lambda k\right) + \mu(\Phi_{X_0},\mathscr{N}_\Pi)\|\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi}\| \end{aligned} \tag{34}$$

$$\begin{aligned}&\|\epsilon_{\Omega_{A_0}}\| = \|\mathscr{P}_{\Omega_{A_0}}(\epsilon_{\Phi_{X_0}} + \mathbf{U}_0\mathbf{V}_0' + \mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi})\| \\ &\leq \mu(\Phi_{X_0},\Omega_{A_0})\left(\|\epsilon_{\Phi_{X_0}}\| + 1\right) + \mu(\mathscr{N}_R,\Omega_{A_0})\|\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Omega}\| \end{aligned} \tag{35}$$

and

$$\begin{aligned}&\|\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Omega}\| \\ &= \|\mathscr{P}_{\mathscr{N}_R\cap\mathscr{N}_\Omega}(\epsilon_{\Phi_{X_0}} + \mathbf{U}_0\mathbf{V}_0' + \epsilon_{\Omega_{A_0}} + \lambda\mathrm{sign}(\mathbf{A}_0))\| \\ &\leq \mu(\Phi_{X_0},\mathscr{N}_\Omega)\big(\|\epsilon_{\Phi_{X_0}}\| + 1\big) \\ &\qquad + \mu(\mathscr{N}_R,\Omega_{A_0}\cap\mathscr{N}_\Omega)\big(\|\epsilon_{\Omega_{A_0}}\| + \lambda k\big). \end{aligned} \tag{36}$$

For convenience introduce the notations $\alpha := \mu(\Phi_{X_0},\Omega_{A_0})$, $\beta := \mu(\mathscr{N}_R,\Omega_{A_0})$, $\xi := \mu(\Phi_{X_0},\mathscr{N}_\Pi)$, and $\nu := \mu(\mathscr{N}_R,\Omega_{A_0}\cap\mathscr{N}_\Pi)$. Then, after mixing (34)–(36) and doing some algebra it follows that

$$\|\Gamma_{\mathscr{N}_R\cap\mathscr{N}_\Omega}\| \leq \theta := \frac{\xi + \lambda k\nu + \alpha(\xi+\alpha\nu)(1-\alpha^2)(\alpha+\lambda k)}{1-\nu\beta-(\xi+\alpha\nu)(1-\alpha^2)(\xi+\alpha\beta)} \tag{37}$$

and

$$\begin{aligned}\|\epsilon_{\Omega_{A_0}}\| \leq&\alpha + (1-\alpha^2)\alpha^2(\alpha+\lambda k) \\ &+ \left[\beta + \alpha^2(1-\alpha^2)\beta + \alpha\xi(1-\alpha^2)^{-1}\right]\theta \end{aligned} \tag{38}$$

$$\|\epsilon_{\Phi_{X_0}}\| \leq (1-\alpha^2)\big[\alpha(\alpha+\lambda k) + (\alpha\beta+\xi)\theta\big]. \tag{39}$$

At this point, it is important to recognize from (12) that $\|\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi}\|_\infty \leq \tau\|\mathbf{\Gamma}_{\mathscr{N}_R\cap\mathscr{N}_\Pi}\|$.

Now building on (37)-(39), one can bound the terms in the r.h.s. of (29) and (30) from above in terms of $\{\alpha,\beta,\xi,\nu,k\}$. Finally, to fulfill C4 and C5, it suffices to confine their corresponding upper bounds to the values 1 and λ, respectively. This imposes the conditions

(a) $\lambda k+\alpha+\alpha(1-\alpha^2)\big[\alpha(\alpha+\lambda k)+(\alpha\beta+\xi)\theta\big]+(1+\nu)\theta < 1$
(b) $\gamma + \eta\alpha\lambda k + (\tau+\eta\alpha+\eta\xi)\theta < \lambda$.

The conditions (a) and (b) imply that C1–C5 hold for the dual certificate $\mathbf{\Gamma}$ if there exists a valid $\lambda \in [\lambda_{\min},\lambda_{\max}]$, with $\lambda_{\max} \geq \lambda_{\min} \geq 0$. The resulting condition is then summarized in the assumptions (I) and (II) of Theorem 1, and the proof is now complete.

REFERENCES

[1] X. Wu, K. Yu, , and X. Wang, "On the growth of Internet application flows: A complex network perspective," in *Proc. IEEE Intl. Conf. on Computer Commun.*, Shangai, China, 2011.

[2] Y. Shavitt, X. Sun, A. Wool, and B. Yener, "Computing the unmeasured: An algebraic approach to Internet mapping," in *Proc. IEEE Intl. Conf. on Computer Commun.*, Alaska, USA, April 2001.

[3] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft, "Structural analysis of network traffic flows," in *Proc. of ACM SIGMETRICS*, New York, NY, Jul. 2004.

[4] P. Barford and D. Plonka, "Characteristics of network traffic flow anomalies," in *Proc. 1st ACM SIGCOMM Workshop on Internet Measurements*, San Francisco, CA, november 2001.

[5] K. Papagiannaki, R. Cruz, and C. Diot, "Network performance monitoring at small time scales," in *Proc. 1st ACM SIGCOMM Workshop on Internet Measurements*, Miami Beach, Florida, october 2003.

[6] E. D. Kolaczyk, *Analysis of Network Data: Methods and Models*. New York: Springer, 2009.

[7] Q. Zhao, Z. Ge, J. Wang, and J. Xu, "Robust traffic matrix estimation with imperfect information: Making use of multiple data sources," vol. 34, pp. 133–144, 2006.

[8] E. Cascetta, "Estimation of trip matrices from traffic counts and survey data: A generalized least-squares estimator," *Transportation Research, Part B: Methodological*, vol. 18, pp. 289–299, 1984.

[9] Y. Vardi, "Network tomography: Estimating source-destination traffic intensities from link data," *Journal of American Statistical Association*, vol. 91, pp. 365 – 377, 1996.

[10] H. V. Zuylen and L. Willumsen, "The most likely trip matrix estimated from traffic counts," *Transportation Research, Part B: Methodological*, vol. 14, pp. 281–293, 1980.
[11] M. Roughan, Y. Zhang, W. Willinger, and L. Qiu, "Spatio-temporal compressive sensing and Internet traffic matrices," *IEEE/ACM Trans. Networking*, vol. 20, pp. 662–676, 2012.
[12] M. Mardani, G. Mateos, and G. B. Giannakis, "Recovery of low-rank plus compressed sparse matrices with application to unveiling traffic anomalie," *IEEE Trans. Info. Theory.*, vol. 59, pp. 5186–5205, Aug 2013.
[13] Y. Zhang, Z. Ge, A. Greenberg, and M. Roughan, "Network anomography," in *Proc. of ACM SIGCOM Conf. on Interent Measurements*, Berekly, CA, USA, Oct. 2005.
[14] V. Chandrasekaran, S. Sanghavi, P. R. Parrilo, and A. S. Willsky, "Rank-sparsity incoherence for matrix decomposition," *SIAM J. Optim.*, vol. 21, no. 2, pp. 572–596, 2011.
[15] M. Mardani, G. Mateos, and G. B. Giannakis, "In-network sparsity regularized rank minimization: Applications and algorithms," *IEEE Trans. Signal Process.*, vol. 59, pp. 5374–5388, Nov. 2013.
[16] ——, "Dynamic anomalography: tracking network anomalies via sparsity and low rank," *IEEE J. Sel. Topics in Signal Process.*, vol. 7, no. 11, pp. 50–66, Feb. 2013.
[17] [Online]. Available: http://internet2.edu/observatory/archive/data-collections.html
[18] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM J. Comput.*, vol. 24, pp. 227–234, 1995.
[19] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM Rev.*, vol. 52, no. 3, pp. 471–501, 2010.
[20] E. J. Candès and T. Tao, "Decoding by linear programming," *IEEE Trans. Info. Theory*, vol. 51, no. 12, pp. 4203–4215, 2005.
[21] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, pp. 925–936, 2009.
[22] A. Ganesh, K. Min, J. Wright, and Y. Ma, "Principal component pursuit with reduced linear measurements," *arXiv:1202.6445v1 [cs.IT]*, 2012.
[23] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.
[24] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1985.
[25] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *Journal of the ACM*, vol. 58, no. 1, pp. 1–37, 2011.
[26] F. Deutsch, *Best Approximation in Inner Product Spaces*, 2nd ed. Springer-Verlag, 2001.
[27] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717–722, 2009.
[28] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, 2nd ed. Athena-Scientific, 1999.
[29] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learning*, vol. 3, pp. 1–122, 2010.
[30] N. Srebro and A. Shraibman, "Rank, trace-norm and max-norm," in *Proc. of Learning Theory*, 2005, pp. 545–560.
[31] S. G. Samko, A. A. Kilbas, and O. I. Marichev, *Fractional Integrals and Derivatives*. Yverdon, Switzerland: Gordon and Breach: Springer, 1993.
[32] T. Eltoft, T. Kim, and T.-W. Lee, "On the multivariate Laplace distribution," *IEEE Signal Process. Letters*, vol. 13, pp. 300–303, May 2006.
[33] J. A. Bazerque, G. Mateos, and G. B. Giannakis, "Rank regularization and Bayesian inference for tensor completion and extrapolation," *IEEE Trans. Signal Process.*, vol. 61, no. 22, pp. 5689–5703, Nov. 2013.
[34] G. B. Giannakis, *Cyclostationary Signal Analysis*. Chapter in Digital Signal Processing Handbook: V. K. Madisetti and D. Williams, Eds. Boca Raton, FL: CRC, 1998.
[35] V. Solo and X. Kong, *Adaptive signal processing algorithms: stability and performance*. Prentice-Hall, 1995.
[36] M. Razaviyayn, M. Hong, and Z.-Q. Luo, "A unified convergence analysis of block successive minimization methods for nonsmooth optimization," *SIAM Journal on Optimization*, vol. 23, pp. 1126–1153, 2013.
[37] Y. Nesterov, "A method of solving a convex programming problem with convergence rate $o(1/k^2)$," *Soviet Mathematics Doklady*, vol. 27, pp. 372–376, 1983.
[38] A. Lakhina, M. Crovella, and C. Diot, "Mining anomalies using traffic feature distributions," vol. 35, pp. 217–228, august 2005.
[39] M. Thottan and C. Ji, "Anomaly detection in IP networks," *IEEE Trans. Signal Process.*, vol. 51, pp. 2191–2204, Aug. 2003.
[40] A. Lakhina, M. Crovella, and C. Diot, "Diagnosing network-wide traffic anomalies," in *Proc. of ACM SIGCOMM*, Portland, OR, Aug. 2004.
[41] M. Mardani, G. Mateos, and G. B. Giannakis, "Subspace learning and imputation for streaming big data matrices and tensors," *arXiv:1404.4667v1 [stat.ML]*, 2014.
[42] H. Kim, S. Lee, X. Ma, and C. Wang, "Higher-order PCA for anomaly detection in large-scale networks," in *Proc. of 3rd Workshop on Comp. Advances in Multi-Sensor Adaptive Proc.*, Aruba, Dutch Antilles, Dec. 2009.
[43] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.