This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2015.2471755, IEEE Transactions on Automatic Control

Robust Distributed Consensus using Total Variation

Walid Ben-Ameur, Pascal Bianchi and Jérémie Jakubowicz

***Abstract*— Consider a connected network of agents endowed with local cost functions representing private objectives. Agents seek to find an agreement on some minimizer of the aggregate cost, by means of repeated communications between neighbors. Consensus on the average over the network, usually addressed by gossip algorithms, is a special instance of this problem, corresponding to quadratic private objectives. Consensus on the median, or more generally, consensus on a given quantile, is also a special instance of this problem. In this paper we show that optimizing the aggregate cost function regularized by a total variation (TV) term has appealing properties. First, it can be done very naturally in a distributed way, yielding algorithms that are efficient on numerical simulations. Secondly, the optimum for the regularized cost is shown to be also the optimum for the initial aggregate cost function under assumptions that are simple to state. Finally, these algorithms are robust to unreliable agents that keep injecting some false value in the network. This is remarkable enough, and is not the case, for instance, of gossip algorithms that are entirely ruled by unreliable agents as detailed in the paper.**

I. Introduction

Total variation (TV) has been widely used in the framework of inverse problems, where the aim is to recover a mathematical object that shows good regularity properties. An important landmark is [1] that successfully applied total variation regularization to image denoising. In the previous decade the role of the L_1 norm was clearly connected to sparsity [2], [3], [4]. In this light, total variation promotes sparsity of the gradient; yielding locally flat objects. Curiously enough, total variation has been mainly developed in the continuous setting, where "bounded variation" functions and their properties are well studied [5]; while in the discrete setting, its properties on graphs, have only been recently emphasized [6], [7]. In this work, we show that total variation regularization can also be useful in the context of consensus problems.

Consensus algorithms designate a class of distributed methods allowing a set of connected agents / nodes to find an agreement on some global parameter value [8]. The latter parameter is often defined as a minimizer of a global objective function defined as the sum of some local regret functions held by the agents [9], [10], [11]. As we shall see below, an important special case is obtained when the aim is to compute the average over the network of some local values held by the agents. The latter scenario will be refered to as the *average consensus* case. It has been well-studied in the literature [8], [12], [13]. The most widespread approach to achieve average consensus is through iteration of linear operations mimicking the behaviour of the heat equation [13]: at each round, nodes average the values in their neighborhood (including themselves). Similarly, in the more general framework of distributed optimization, many algorithms have been proposed: some of them are based on distributed (sub)gradient approaches [10], [11], [14], [15] while others use splitting methods such as the Alternating Direction Method of Multipliers (ADMM) (see [16], [17], [18] and references therein). Under certain hypotheses, such approaches can be shown to converge to a state where each node in the network eventually has the same value – the sought parameter.

However, most of these works share a common view of the network: all agents show good will. They do not, for instance, deliberately introduce some false value inside the network, or refuse to update their value. There are a few recent work raising the problem of misbehaving agents in the gossip process [19], [20], see also [21] for a general perspective. In such scenarios, standard consensus algorithms not only fail, but can be driven arbitrarily far away from the sought consensus [22]. A first approach to increase consensus robustness in unreliable networks is to detect misbehaving agents, identify them and finally exclude them from the network. Of course, cleaning the network beforehand is certainly beneficial whenever feasible, however misbehaving agents are not necessarily detectable and even if they are, may be detectable only by using involved and computationally expensive algorithms. We refer to the recent works [23], [24]. An alternative is to design simple algorithms that naturally show good robustness properties. For instance, the authors of [25] study a continuous-time dynamical system allowing agents to track an agreement in the presence of external perturbation. The authors of [26] introduce a scheme in which each agent combines her/his current estimate with all but the extremal values received from her/his neighborhood. In the above works, it is worth mentioning that the objective is to ensure an (approximate) agreement between normal agents, irrespective to the value of this agreement. In this paper, our main interest is to build a robust consensus algorithm allowing the agents to find an agreement on a *sought* parameter value.

Contribution. Our contributions are the following. First, our definition of total variation on a graph is distinct from previous works [6], [7]. The distinction is that we use what is known as the *anisotropic* total variation in the context of images and meshes. This subtle distinction has important consequences as it allows simple distributed algorithms.

Second, we cast the problem of robust distributed optimization over a network as an inverse problem using total variation regularization. To the best of our knowledge, this is a new usecase for total variation regularization. Loosely speaking, this viewpoint amounts to thinking of consensus in a network as an extreme denoising process, where all the agents share the same value. In the context of image processing, it amounts to denoising until the image becomes totally grey, meaning that all pixels in the image would share the same grey level. See also [27] for a regularization approach to consensus networks.

Third, using our TV framework, we provide closed-form sufficient regularity conditions under which the minimizers of the relaxed problem coincide with the sought minimizers of the initial optimization problem. This is the main contribution of this paper. As a sanity check for the robustness of our algorithms, we analyze the convergence of our algorithms in the presence of stubborn agents that permanently introduce some false value in the network. We prove that unlike traditional approaches, our algorithms ensure that the estimates cannot be driven arbitrarily far away from the sought consensus.

Last, we provide two iterative distributed algorithms which are shown to converge to the minimizers of the relaxed problem. Experimentally, we observe good convergence properties for the second (ADMM-based) algorithm.

The paper is organized as follows. Section II introduces the problem. Section V provides preliminary material on discrete Total Variation. Section III is devoted to the study of the minimizers of a relaxed distributed optimization problem . Algorithms are proposed in Section IV. In Section VI, we analyze the convergence of our algorithms in a scenario where stubborn agents are present. Section VII presents the numerical results. Some of the proofs are given in appendixes.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2015.2471755, IEEE Transactions on Automatic Control

II. Distributed Optimization

A. Notations

1) Generalities: The set of *integers* is denoted $\mathbb{N}$. The set of *real numbers* is denoted $\mathbb{R}$. The cardinality of a finite set S is denoted $|S|$. Given two sets S and T, the set of *functions* from S to T is denoted T^S. In other terms an *element* f of T^S is a function $f : S \to T$, with *domain* S and *codomain* T. Function composition is denoted by $\circ$; *i.e.*, $f \circ g$ stands for the function $x \mapsto f(g(x))$, provided the domain of f contains the codomain of g. Vector spaces are denoted with bold uppercase symbols. For instance, $\mathbb{R}^S$, equipped with standard scalar operations, is a vector space. Vectors are denoted with bold lowercase symbols. When S is a finite set, $\mathbb{R}^S$ equipped with dot product $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{s \in S} \boldsymbol{x}(s)\boldsymbol{y}(s)$ is a Euclidean vector space. Constant function $s \in S \mapsto 0 \in \mathbb{R}$ (resp. $s \in S \mapsto 1 \in \mathbb{R}$) is denoted $\mathbf{0}_S$ (resp. $\mathbf{1}_S$), or simply $\mathbf{0}$ (resp. $\mathbf{1}$), when no confusion can occur. Given $\boldsymbol{x} \in \mathbb{R}^S$ we use notation $\bar{\boldsymbol{x}}$ for $\frac{1}{|S|}\langle \boldsymbol{x}, \mathbf{1} \rangle$. Notice that $\bar{\boldsymbol{x}}$ denotes a scalar. The subspace $\{\lambda \mathbf{1}_S; \lambda \in \mathbb{R}\} \subset \mathbb{R}^S$ has dimension 1 and is referred to as the *consensus subspace* and denoted $\boldsymbol{C}_S$, or simply $\boldsymbol{C}$, when S is clear from the context. For a subset $T \subset S$, $\mathbf{1}_T$ denotes the function that take value 1 on T and 0 on $S \backslash T$.

2) Graphs: We will consider undirected finite graphs $G = (V, E)$ where set V denotes the *vertex* set and set E consists of *edges* $\{v, w\}$. We will also use notation $v \sim w$ to express the fact that vertices v and w are *neighbors* in G, *i.e.* $\{v, w\} \in E$. We denote the *degree* of a vertex v by $d(v)$. All the graphs we will consider are assumed to be *connected*. For a subset $A \subset V$, its *perimeter* is expressed as

$$\mathrm{Per}(A) = |\{v \sim w : v \in A, w \notin A\}|$$

In other terms, the perimeter of a set is the number of edges from that set to "outside" that set.

3) Total variation: Given a graph $G = (V, E)$ and a function $\boldsymbol{x} \in \mathbb{R}^V$, we denote by $\|\boldsymbol{x}\|_{\mathrm{TV}}$ the quantity $\sum_{\{v,w\} \in E} |\boldsymbol{x}(v) - \boldsymbol{x}(w)|$. Note that $\| \cdot \|_{\mathrm{TV}}$ is *not* a norm. It is a *seminorm* on $\mathbb{R}^V$. Since G is assumed to be connected, it can be seen that $\| \cdot \|$ is a norm on the subspace $\mathbb{R}^V_0$ of $\mathbb{R}^V$ defined as $\{\boldsymbol{x} \in \mathbb{R}^V : \langle \boldsymbol{x}, \mathbf{1} \rangle = 0\}$. Notice that $\mathrm{Per}(A) = \|\mathbf{1}_A\|_{\mathrm{TV}}$. In the more general case where $\boldsymbol{x}(v)$ is itself a vector in a vector space equipped with a norm $\| \cdot \|$, total variation is defined as $\sum_{\{v,w\} \in E} \|\boldsymbol{x}(v) - \boldsymbol{x}(w)\|$. In this case, total variation obviously depends on the norm $\| \cdot \|$.

B. The Problem

Consider a network of agents represented by an undirected finite graph $G = (V, E)$, we investigate the following optimization problem:

$$\inf_{x \in \mathbb{R}} \sum_{v \in V} f_v(x) \tag{1}$$

where $f_v : \mathbb{R} \to \mathbb{R}$ is a function which can be interpreted as the regret of agent v when agent v holds value x. Merely for notational convenience, this paper is restricted to the case where parameter x is real. Generalization to the case where x belongs to an arbitrary Euclidean space is however straightforward. We assume the following.

Assumption 1:
(a) For any $v \in V$, f_v is a convex function.
(b) The infimum of (1) is attained at some point $x \in \mathbb{R}$.

Example 1: We shall pay special attention to the following particular case, which we shall refer to as the *Average Consensus (AC)* case:

$$\text{(AC)} \quad f_v(x) = \frac{1}{2}(x - \boldsymbol{x}_0(v))^2 \tag{2}$$

where $\boldsymbol{x}_0(v)$ represents some initial value held by agent v. In that case, problem (1) is equivalent to the distributed computation of the average $\bar{\boldsymbol{x}}_0 = (1/|V|) \sum_v \boldsymbol{x}_0(v)$.

Example 2: A second special case of interest will be refered to as the *Median Consensus (MC)* case:

$$f_v(x) = |x - \boldsymbol{x}_0(v)| \, . \tag{3}$$

In this scenario, solving problem (1) is equivalent to searching for the median of sequence $(\boldsymbol{x}_0(v))_{v \in V}$.
Notice that other choices of functions f_v could be mentioned such as the so called *quantile loss* $f_v(x) = \tau(x - \boldsymbol{x}_0(v))_+ - (1-\tau)(x - \boldsymbol{x}_0(v))_-$ that would yield more general quantiles than the median. Notice also that, with the definition provided, there could well be infinitely many medians. For instance, a network with a couple of agents having initial values 1 and 2 respectively, would admit the whole interval $[1, 2]$ as admissible medians. This is an expected consequence of the fact that the associated problem is not strictly convex.

Each agent v is supposed to hold some value $\boldsymbol{x}_n(v)$ at each time $n \in \mathbb{N}$. The aim of this paper is to introduce and analyze distributed algorithms which, under some assumptions, drive all sequences $(\boldsymbol{x}_n(v))_{v \in V}$ to a common minimizer of (1) as n tends to infinity. Moreover, the proposed algorithms should be robust to the presence of misbehaving agents. By robust, we mean that the final estimate of regular (well-behaved) agents should remain in an acceptable vicinity of the sought consensus even in the case when other agents permanently introduce some false value in the network.

C. Network Model

Throughout this paper, we assume a synchronous network where a global clock allows the agents to communicate with each other at each clock tick. In this paper, we refer to a *distributed algorithm* as an iteration of the form:

$$\boldsymbol{x}_{n+1}(v) = h_{n,v}\left((\boldsymbol{x}_k(v), \boldsymbol{x}_k(w) : w \sim v, 0 \le k \le n)\right) \tag{4}$$

for some specified functions $h_{n,v}$. Nevertheless, we shall sometimes assume that some subset $S \subset V$ of agents do not follow the specified update rule (4). Such agents will be called *irregular*. An irregular agent $v \in S$ is called *stubborn* if for any $n \ge 0$,

$$\boldsymbol{x}_n(v) = \boldsymbol{x}_0(v) \, . \tag{5}$$

We denote by S the set of irregular agents and by $R = V \backslash S$ the set of *regular* agents.

D. Variational Framework

Consider replacing problem (1) with:

$$\min_{\boldsymbol{x} \in \boldsymbol{C}} \sum_{v \in V} f_v(\boldsymbol{x}(v)) \tag{6}$$

Both problems are indeed equivalent using the one-to-one correspondence:

$$x \in \mathbb{R} \leftrightarrow \boldsymbol{x} = x\mathbf{1}_V \in \boldsymbol{C}$$

Considering the indicator function of $\boldsymbol{C}$, defined as $\iota_{\boldsymbol{C}}(\boldsymbol{x}) = 0$ if $\boldsymbol{x} \in \boldsymbol{C}$ and $\iota_{\boldsymbol{C}}(\boldsymbol{x}) = +\infty$ otherwise, problem (6) leads to:

$$\min_{\boldsymbol{x} \in \mathbb{R}^V} \sum_{v \in V} f_v(\boldsymbol{x}(v)) + \iota_{\boldsymbol{C}}(\boldsymbol{x}) \tag{7}$$

Both problems (6) and (7) are straightforwardly equivalent. In what follows, we will study the impact of replacing problem (7) with a distinct problem (8):

$$\min_{\boldsymbol{x} \in \mathbb{R}^V} \sum_{v \in V} f_v(\boldsymbol{x}(v)) + \lambda \|\boldsymbol{x}\|_{\mathrm{TV}} \tag{8}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2015.2471755, IEEE Transactions on Automatic Control

Both problems (7) and (8) can be written:

$$\min_{x\in\mathbb{R}^V} \sum_{v\in V} f_v(\boldsymbol{x}(v)) + U(\boldsymbol{x}) \tag{9}$$

where $U : \mathbb{R}^V \to \mathbb{R}$ is a convex regularization penalizing the functions $\boldsymbol{x} \in \mathbb{R}^V$ that are away from the consensus subspace $\boldsymbol{C}$. At an intuitive level, setting $U = \iota_{\boldsymbol{C}}$ means that consensus must be achieved *at any price*, while setting $U = \lambda\|\cdot\|_{\mathrm{TV}}$ softly drives solutions to $\boldsymbol{C}$, depending on how large λ is. Using this viewpoint, problem (8) is a relaxation of problem (7), in the sense that the solutions do not necessarily belong to $\boldsymbol{C}$. In the presence of irregular agents, it could be beneficial to break the diktat of consensus, to allow regular agents to possibly disagree with irregular ones. Of course, it can then no longer be expected that the minimizers of (8) coincide in full generality with those of (7). Yet, as we are going to see in the sequel (see Theorem 1), there exist concise quantitative conditions on λ and f_v guaranteeing that minimizers of (8) and (7) will coincide exactly. In the AC problem, functions f_v are given by (2) and the problem (8) reduces to:

$$\min_{x\in\mathbb{R}^V} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{x}_0\|_2^2 + \lambda\,\|\boldsymbol{x}\|_{\mathrm{TV}}\ . \tag{10}$$

In the context of image processing, the particular objective function (10) is referred to as the ROF (Rudin-Osher-Fatemi) energy [1]. We will refer to the general objective function in (8) as a regularized energy, and to the minimizers of (8) as regularized minimizers.

Our aim is threefolds: *i)* to prove that the minimizers of (1) coincide with the regularized minimizers at least for a specified class of functions f_v; *ii)* to propose distributed algorithms to find regularized minimizers, *iii)* to quantify the robustness of the algorithms in the presence of irregular (stubborn) agents.

III. Regularized minimizers

In this section we provide quantitative results linking total variation regularization and consensus. The general idea being that, in the same way $\|\cdot\|_1$ norm regularization yields sparse vectors, $\|\cdot\|_{\mathrm{TV}}$ regularization yields to flat functions, *i.e.* piecewise constant functions. Consensus corresponds to globally constant functions, *i.e.* extremely flat functions, attained for sufficiently strong regularization term. How large should the regularization parameter λ be set to, in order to reach consensus, is intimately linked to the dual norm associated with total variation which we now present.

A. Total Variation Dual Norm

Recall that $\mathbb{R}_0^V$ denotes the subspace $\{\boldsymbol{x}\in\mathbb{R}^V : \langle \boldsymbol{x}, \mathbf{1}_V\rangle = 0\} \subset \mathbb{R}^V$ of zero-mean functions over V; and that $\|\cdot\|_{\mathrm{TV}}$ is a norm on $\mathbb{R}_0^V$ since G is assumed connected. The dual space $(\mathbb{R}_0^V)^*$ identified with $\mathbb{R}_0^V$ using the standard scalar product is equipped with the dual norm:

$$\|\boldsymbol{u}\|_* = \max_{\|\boldsymbol{x}\|_{\mathrm{TV}}\le 1, \boldsymbol{x}\in\mathbb{R}_0^V} \langle \boldsymbol{x}, \boldsymbol{u}\rangle\,. \tag{11}$$

In the case where $\boldsymbol{x}(v)$ is itself a vector in a vector space equipped with a norm $\|\cdot\|$, this definition holds with the convention that $\langle \boldsymbol{x}, \boldsymbol{u}\rangle = \sum_v \langle \boldsymbol{x}(v), \boldsymbol{u}(v)\rangle$ and depends on the norm $\|\cdot\|$. We introduce the unit ball for the dual norm $\|\cdot\|_*$:

$$B_* = \{\boldsymbol{u} \,:\, \|\boldsymbol{u}\|_* \le 1\}\ .$$

The following property is a well-known consequence of a general fact about subdifferentials of support functions (see for instance Proposition 3.2.1 [28, p.221]):

Proposition 1: If $\partial\|\boldsymbol{x}\|_{\mathrm{TV}}$ denotes the subdifferential of norm $\|\cdot\|_{\mathrm{TV}}$ at point $\boldsymbol{x}$, one has:

$$\partial\|\boldsymbol{x}\|_{\mathrm{TV}} = \{\boldsymbol{u}\in\mathbb{R}_0^V : \|\boldsymbol{u}\|_* \le 1, \langle \boldsymbol{u}, \boldsymbol{x}\rangle = \|\boldsymbol{x}\|_{\mathrm{TV}}\}$$

In particular, $\partial\|0\|_{\mathrm{TV}} = B_*$.

B. Total Variation Regularization And Consensus

Define function $F : \mathbb{R}^V \to \mathbb{R}$ by $F(\boldsymbol{x}) = \sum_v f_v(\boldsymbol{x}(v))$. For any $x \in \mathbb{R}$, one has:

$$\partial F(x\mathbf{1}) = \left\{\boldsymbol{u}\in\mathbb{R}^V : \forall v\in V,\ \boldsymbol{u}(v) \in \partial f_v(x)\right\}\ .$$

When all f_v's are differentiable, note that $\partial F(x\mathbf{1}_V)$ is a singleton $\{(f_v'(x))_{v\in V}\}$. Recall that B_* is the unit ball associated with the dual norm.

Theorem 1: Under Assumption 1, among the following statements, 1), 2) and 3) are equivalent and imply 4).

1) $x^\star\mathbf{1}_V$ is a minimizer of (8) ;
2) $\partial F(x^\star\mathbf{1}_V) \cap \lambda B_*$ is nonempty ;
3) There exists $\boldsymbol{u} \in \partial F(x^\star\mathbf{1}_V)$ such that $\sum_{v\in V}\boldsymbol{u}(v) = 0$ and for all $A \subset V$,
$$\sum_{v\in A}\boldsymbol{u}(v) \le \lambda\,\mathrm{Per}(A)\,.$$
4) $x^\star$ is a minimizer of (1).

Proof: [1) $\Leftrightarrow$ 2)] Note that $x^\star\mathbf{1}_V$ is a minimizer of $F + \lambda\|\cdot\|_{\mathrm{TV}}$ iff $0 \in \partial F(x^\star\mathbf{1}_V) + \lambda\partial\|x^\star\mathbf{1}_V\|_{\mathrm{TV}}$. From Proposition 1, $\partial\|x^\star\mathbf{1}_V\|_{\mathrm{TV}} = B_*$. Therefore, 1) holds iff there exists $\boldsymbol{u} \in \partial F(x^\star\mathbf{1}_V)$ such that $0 \in \boldsymbol{u} + \lambda B_*$. Otherwise stated, there exists $\boldsymbol{u} \in \partial F(x^\star\mathbf{1}_V)$ such that $\boldsymbol{u} \in \lambda B_*$. [2) $\Leftrightarrow$ 3)] is a consequence of Proposition 5. [3) $\Rightarrow$ 4)] As $\sum_v \partial f_v = \partial(\sum_v f_v)$, condition $\sum_{v\in V}\boldsymbol{u}(v) = 0$ implies that $0 \in \partial(\sum_v f_v)(x^\star)$. Thus, $x^\star$ is a minimizer of $\sum_v f_v$. ■

In order to have some insights about Theorem 1, assume for instance that all f_v's are differentiable. Condition 2) of Theorem 1 can be simply rewritten as

$$\|\nabla F(x^\star\mathbf{1}_V)\|_* \le \lambda \tag{12}$$

where $\nabla F(x^\star\mathbf{1}_V)$ is the gradient vector whose vth component is $f_v'(x^\star)$. Now consider a solution $x^\star \in \mathbb{R}$ to the initial problem (1). Theorem 1 states that whenever this solution satisfies (12), then $x^\star\mathbf{1}_V$ is also a solution to the relaxed problem (8). The criterion $\|\nabla F(x^\star\mathbf{1}_V)\|_* \le \lambda$ is complicated to grasp due to the nature of $\|\cdot\|_*$. However it is reminiscent of a very well known fact: $\|\nabla F(\boldsymbol{x})\| = 0$ says that $\boldsymbol{x}$ is critical point. The result here has the same flavour: a point is a critical point for a penalized function when the gradient is small for a well chosen norm. And small relates to λ the quantity of regularization injected. Of course, condition (12) remains a little abstract unless we have a way to verify the latter. Algorithm 3 of Section V-C provides a practical method to compute the dual norm which can be used to verify condition (12) in practice. By statement 3) of Theorem 1, condition (12) is equivalent to:

$$\forall A \subset V,\ \sum_{v\in A} f_v'(x^\star) \le \lambda\,\mathrm{Per}(A)\ . \tag{13}$$

We thus have the following Corollary.

Corollary 1: Assume that f_v is differentiable for all $v \in V$. Let $x^\star$ be a minimizer of (1) satisfying condition (13). Then, $x^\star\mathbf{1}_V$ is a minimizer of (8).

This result reveals that, provided λ be large enough, the critical value being $\|\nabla F(x^*\mathbf{1}_V)\|_*$, consensus is a minimizer of eq. (8). Qualitatively, this result could be derived from the following remark.

Remark 1: One can recognize in eq. (8) an exact penalty method associated with the constraints $\forall v \sim w, \boldsymbol{x}(v) = \boldsymbol{x}(w)$, (see for instance [28, p.298]). Now, it is known ([28, p.321]) that, for λ large enough, the regularized minimizers associated with an exact penalty automatically satisfy the constraints. The aim of Theorem 1 and its corollary is to give a closed-form and concise solution to the critical λ needed for the constraints to be automatically satisfied. Notice that the expression for the critical λ is not provided by the exact penalty theory [28].

We now review the consequences of Theorem 1 regarding the Average Consensus and the Median Consensus problems described earlier in the introduction.

C. Average Consensus Problem

Proposition 2: The following statements are equivalent.

1) $\bar{\boldsymbol{x}}_0 \mathbf{1}_V$ is the unique minimizer of (10) ;
2) $\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0\, \mathbf{1}_V \in \lambda B_*$;
3) For all $A \subset V$,

$$\left| \frac{\sum_{v\in A} \boldsymbol{x}_0(v)}{|A|} - \bar{\boldsymbol{x}}_0 \right| \leq \lambda \frac{\mathrm{Per}(A)}{|A|}$$

Proposition 2 quantifies how much a local average can fluctuate around $\bar{\boldsymbol{x}}_0$ in order to preserve the sought equilibrium at $\bar{\boldsymbol{x}}_0 1$: the larger the ratio $\mathrm{Per}(A)/|A|$ the more it can fluctuate safely inside A. The heuristic behind this argument is that the ratio $\mathrm{Per}(A)/|A|$ measures how well a given region A is connected to the rest of the network, since $\mathrm{Per}(A)$ is the number of edges connecting A to its complementary set $V \backslash A$ and $|A|$ is a measure of its size. A large ratio $\mathrm{Per}(A)/|A|$ amounts to saying that relatively to its size, A has a lot of connections to its outside. Quantity $\mathrm{Per}(A)/|A|$ is a well studied in the literature and is linked with so called Cheeger isoperimetric constant [29].

For $\lambda > \|\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0 \mathbf{1}\|_*$, the minimizer of (10) is *exactly* $\bar{\boldsymbol{x}}_0 \mathbf{1}$. The critical value $\|\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0 \mathbf{1}\|_*$ can be computed using algorithm 3 provided in Section V-C. However, the following result can be useful to set a crude λ in the range $[\|\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0 \mathbf{1}\|_*, +\infty)$, without any additional computational burden.

Corollary 2: Assume that $\forall v \in V$, $|\boldsymbol{x}_0(v)| \leq M$. Then,

$$\|\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0 \mathbf{1}\|_* \leq \frac{M|V|}{2}$$

Proof: For all $A \subset V$, such that $|A| \leq |V|/2$, one has: $\|\mathbf{1}_A\|_{\mathrm{TV}}^{-1} \langle \mathbf{1}_A, \boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0 \mathbf{1} \rangle \leq \frac{|A|}{\mathrm{Per}(A)} M \leq \frac{|V|M}{2}$. Now, from proposition 5:

$$\|u\|_* = \max_{A\subset V, |A|\leq |V|/2} \langle u, \mathbf{1}_A / \|\mathbf{1}_A\|_{\mathrm{TV}} \rangle$$

which concludes the proof. ■

D. Median Consensus Problem

We denote by $\mathrm{median}(\boldsymbol{x}_0)$ the set of minimizers of (1) when $f_v(x) = |x - \boldsymbol{x}_0(v)|$ for all $v \in V$. It is straightforward to show that

$$\begin{array}{lll} \mathrm{median}(\boldsymbol{x}_0) = & \{\boldsymbol{x}_0 \circ \sigma(\frac{|V|+1}{2})\} & \text{if } |V| \text{ is odd} \\ \mathrm{median}(\boldsymbol{x}_0) = & \left[\boldsymbol{x}_0 \circ \sigma(\frac{|V|}{2}) \,,\, \boldsymbol{x}_0 \circ \sigma(\frac{|V|}{2} + 1)\right] & \text{if } |V| \text{ is even.} \end{array}$$

where $\sigma : \{1, \ldots, |V|\} \to V$ is any bijection such that $(\boldsymbol{x}_0 \circ \sigma)(1) \leq \cdots \leq (\boldsymbol{x}_0 \circ \sigma)(|V|)$. We use notation $\mathrm{median}(\boldsymbol{x}_0)\mathbf{1}_V$ to designate the set of functions $\{x\mathbf{1}_V : x \in \mathrm{median}(\boldsymbol{x}_0)\}$. We introduce the following sequence (formally a function with domain $\{1, \ldots, |V|\}$ and codomain $\mathbb{R}$):

$$\boldsymbol{d} = \begin{cases} (-1, \cdots, -1, 0, 1, \cdots 1) & \text{if } |V| \text{ is odd} \\ (-1, \cdots, -1, 1, \cdots 1) & \text{if } |V| \text{ is even.} \end{cases}$$

We define by $\mathcal{U}$ the set of functions $\boldsymbol{u} \in \mathbb{R}^V$ for which there exists a bijection $\sigma : \{1 \cdots |V|\} \to V$ such that $\boldsymbol{u} \circ \sigma = \boldsymbol{d}$. Otherwise stated, $\mathcal{U}$ is the set of all permutations of sequence $\boldsymbol{d}$. We set:

$$\lambda_0 = \max\{\|\boldsymbol{u}\|_* : \boldsymbol{u} \in \mathcal{U}\} \,.$$

Proposition 3: For any $\lambda > \lambda_0$, the set of regularized minimizers of the (MC) problem is equal to $\mathrm{median}(\boldsymbol{x}_0)\mathbf{1}_V$.

One might also consider using the following cruder bound.

Corollary 3: For any $\lambda > |V|/2$, the set of regularized minimizers of the (MC) problem is equal to $\mathrm{median}(\boldsymbol{x}_0)\mathbf{1}_V$.

Proof: Remark that, for $\varphi : x \in \mathbb{R} \mapsto |x - a| \in \mathbb{R}$ and any $y \in \partial\varphi(x)$, one has: $|y| \leq 1$. Now, for any vector $\boldsymbol{u}$ such that $\forall v \in V$, $|\boldsymbol{u}(v)| \leq 1$ and $\sum_{v\in V} \boldsymbol{u}(v) = 0$, the following holds:

$$\|\boldsymbol{u}\|_* = \max_{A\subset V, |A|\leq |V|/2} \langle \boldsymbol{u}, \|\mathbf{1}_A\|_{\mathrm{TV}}^{-1} \mathbf{1}_A \rangle \leq \frac{|A|}{\mathrm{Per}(A)} \leq \frac{|V|}{2}$$

■

IV. Proposed Algorithms

In this section we present two *distributed* algorithms that aim at solving problem (8).

A. Subgradient Algorithm

Note that function $\lambda\|\boldsymbol{x}\|_{\mathrm{TV}}$ is non-differentiable. Perhaps the most simple and natural approach is to use the subgradient algorithm associated with problem (8). This naturally yields the following distributed algorithm, where each node v holds an estimate $\boldsymbol{x}_n(v)$ of the minimizer at time n and combine it with the ones received from its neighbors.

Algorithm 1 Subgradient Descent

Subgradient Descent

Input: $\boldsymbol{x}_0 \in \mathbb{R}^V$, $\lambda > 0$, $(\gamma_n)_{n>0}$
$n \leftarrow 1$
repeat
  for all $v \in V$ **do**
    Compute Subgradient $\boldsymbol{g}_n(v) \in -\partial f_v(\boldsymbol{x}_n(v))$
    Update using:

$$\boldsymbol{x}_{n+1}(v) = \boldsymbol{x}_n(v) + \gamma_n \left[\boldsymbol{g}_n(v) + \lambda \sum_{w\sim v} \mathrm{sign}\left(\boldsymbol{x}_n(w) - \boldsymbol{x}_n(v)\right) \right]$$

    $n \leftarrow n + 1$
  end for
until Termination Condition

Standard convex optimization arguments can be used to prove that function $\boldsymbol{x}_n$ converges to a minimizer of (9) under the hypothesis of decreasing step size. The arguments being standard, the proof is omitted and we refer to [30], [31] or references therein. A simple argument given in the appendix can be used to prove that the following property holds: $\forall n \geq 0, \bar{\boldsymbol{x}}_n = \bar{\boldsymbol{x}}_0$. The following theorem precisely sums up the mentioned results.

Assumption 2: The following holds.

1) The step sizes satisfy $\gamma_n > 0$ for all n, $\sum_n \gamma_n = +\infty$ and $\sum_n \gamma_n^2 < \infty$.
2) There exists a constant $C > 0$ such that for any $v \in V$, any $x \in \mathbb{R}$ any $g \in \partial f_v(x)$, $|g| \leq C(1 + |x|)$.

Notice that point 2) in the previous assumption is automatically satisfied in the case where the functions f_v are differentiable with Lipschitz gradient.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2015.2471755, IEEE Transactions on Automatic Control

Theorem 2: Consider that $R = V$ (all agents are regular). Under Assumptions 1, 2, sequence $\boldsymbol{x}_n$ given by Algorithm IV-A converges to the minimizers of (8). Moreover, in the average consensus case (AC), the following property holds: $\forall n \geq 0, \bar{\boldsymbol{x}}_n = \bar{\boldsymbol{x}}_0$.

Theorem 2 is an immediate consequence of Proposition 8 given in Appendix II-A. This theorem is also a consequence of well known results on the subgradient algorithm [32]. However, for the sake of completeness, we provide a self-contained proof in this paper.

B. Alternating Direction of Method of Multipliers (ADMM)

It is a widely acknowledged fact that the subgradient method is slow in terms of convergence rate (see [33, Chap. 3.2.3]). Many alternatives do exist in order to speed up the convergence [34], [30]. Among these solutions, we propose an approach which can be seen as a special case of ADMM. We refer to [17], [18], [16], [35] for other examples of applications of ADMM to distributed optimization.

Let us denote by $(V, \overleftrightarrow{E})$ the directed graph such that: $(v, w) \in \overleftrightarrow{E}$ iff $v \sim w$. That is, each pair $\{v, w\}$ of connected nodes yields two edges in $\overleftrightarrow{E}$ (one from v to w, the other from w to v). Problem (8) is equivalent to

$$\min_{\substack{(\boldsymbol{x},\boldsymbol{z})\in\mathbb{R}^V\times\mathbb{R}^{\overleftrightarrow{E}}: \\ s.t.\, \boldsymbol{z}(v,w)=\boldsymbol{x}(w), \forall(v,w)\in\overleftrightarrow{E}}} \sum_{v\in V} f_v(\boldsymbol{x}(v)) + \lambda \sum_{\{v,w\}\,:\,v\sim w} |\boldsymbol{z}(v,w) - \boldsymbol{z}(w,v)|$$

The augmented Lagrangian writes:

$$\mathcal{L}(x, z; \eta) = \sum_{v\in V} f_v(\boldsymbol{x}(v)) + \sum_{\{v,w\}\,:\,v\sim w} \lambda |\boldsymbol{z}(v,w) - \boldsymbol{z}(w,v)| + \sum_{(v,w)\in\overleftrightarrow{E}} T_\rho\left(\boldsymbol{\eta}(v,w), \boldsymbol{z}(v,w) - \boldsymbol{x}(w)\right)$$

where we set $T_\rho(\alpha, \beta) = \alpha\beta + \frac{\rho}{2}\beta^2$. The ADMM consists in generating three sequences $(\boldsymbol{x}_n, \boldsymbol{z}_n, \boldsymbol{\eta}_n)_{n\geq 0}$ recursively defined by

$$\begin{aligned}
\boldsymbol{x}_{n+1} &= \arg\min_{\boldsymbol{x}\in\mathbb{R}^V} \mathcal{L}(\boldsymbol{x}, \boldsymbol{z}_n; \boldsymbol{\eta}_n) \\
\boldsymbol{z}_{n+1} &= \arg\min_{\boldsymbol{z}\in\mathbb{R}^{\overleftrightarrow{E}}} \mathcal{L}(\boldsymbol{x}_{n+1}, \boldsymbol{z}; \boldsymbol{\eta}_n) \\
\boldsymbol{\eta}_{n+1}(v,w) &= \boldsymbol{\eta}_n(v,w) + \rho\left(\boldsymbol{z}_{n+1}(v,w) - \boldsymbol{x}_{n+1}(w)\right)
\end{aligned}$$

for all $(v, w) \in \overleftrightarrow{E}$. In Appendix II-B, we make ADMM explicit and prove that the update equation in $\boldsymbol{x}_n$ is given by Algorithm IV-B below.

Denote by $\text{proj}_{[-\omega,\omega]}(x)$ the projection of x onto $[-\omega, \omega]$ and by $\text{prox}_{f,\rho}(x) = \arg\min_y f(y) + \frac{\rho}{2}(y - x)^2$ the proximal operator associated with a real function f.

We refer to [17, p.20] for a discussion of the way to adjust ρ in practice. Notice that Theorem 3 shows that, on a theoretical side, convergence is not affected by a change in ρ.

Notice that $\text{prox}_{f,\rho}(x) = a + \frac{\rho}{1+\rho}(x - a)$ for $f(x) = \frac{1}{2}(x - a)^2$ (average consensus case) and $\text{prox}_{f,\rho}(x) = a + \max(|x - a| - \rho^{-1}, 0)\,\text{sign}(x - a)$ for $f(x) = |x - a|$ (median consensus case).

Here, each regular agent v not only maintains an estimate $\boldsymbol{x}_n(v)$ of the minimizer, but also holds in its memory one scalar $\boldsymbol{\mu}_n(w, v)$ for any of its neighbors $w \sim v$. We stress the fact that the values $\boldsymbol{\mu}_n(w, v)$ are purely private in the sense that they are not exchanged by agents. Otherwise stated, at time n, an agent v only shares her/his estimates $\boldsymbol{x}_n(v)$ with her/his neighbors, the variable $\boldsymbol{\mu}_n(v)$ being private. In the (AC) case, operator $\text{prox}_{f_v,\rho}$ has a simple expression. In that case, the update equation in $\boldsymbol{x}_n$ simplifies to:

$$\boldsymbol{x}_{n+1}(v) = \frac{\boldsymbol{x}_0(v) + \rho\boldsymbol{d}(v)\left(\boldsymbol{x}_n(v) + \frac{3}{2}\tilde{\boldsymbol{\mu}}_{n+1}(v) - \frac{1}{2}\tilde{\boldsymbol{\mu}}_n(v)\right)}{1 + \rho\boldsymbol{d}(v)}.$$

Algorithm 2 ADMM

ADMM
Input: $\boldsymbol{x}_0 \in \mathbb{R}^V$, $\lambda > 0$, $\rho > 0$
Initialization Step
$n \leftarrow 0$
for all $v \in V$ **do**
  Initialize: $\boldsymbol{\mu}_0(v)$
  for all neighbors $w \sim v$ **do**
    Initialize: $\boldsymbol{\mu}_0(w, v)$
  end for
end for
repeat
  $n \leftarrow n + 1$
  for all $v \in V$ **do**
    for all neighbors $w \sim v$ **do**
      Update

$$\boldsymbol{\mu}_n(w,v) = \text{proj}_{[-2\lambda/\rho, 2\lambda/\rho]}\left(\boldsymbol{\mu}_{n-1}(w,v) + \boldsymbol{x}_{n-1}(w) - \boldsymbol{x}_{n-1}(v)\right)$$

    end for
    Compute $\tilde{\boldsymbol{\mu}}_n(v) = \frac{1}{d(v)} \sum_{w\sim v} \boldsymbol{\mu}_n(w, v)$
    Update

$$\boldsymbol{x}_n(v) = \text{prox}_{f_v,\,\rho d(v)}\left(\boldsymbol{x}_{n-1}(v) + \tfrac{3}{2}\tilde{\boldsymbol{\mu}}_n(v) - \tfrac{1}{2}\tilde{\boldsymbol{\mu}}_{n-1}(v)\right)$$

  end for
until Termination Condition

As Algorithm IV-B can be seen as a special case of a standard ADMM, the following result follows directly from [34].

Theorem 3: Assume that $R = V$ (all agents are regular). Under Assumption 1, sequence $\boldsymbol{x}_n$ given by Algorithm IV-B converges to the minimizers of (8). Moreover, in the average consensus case (AC), if the graph G is assumed d-regular, *i.e.* each node has the same degree d, the average is preserved over the network: $\forall n \geq 0, \bar{\boldsymbol{x}}_n = \bar{\boldsymbol{x}}_0$.

V. Dual Norm Computation

In the previous sections we detailed how the critical amount of regularization needed for consensus can be expressed in terms of the TV dual norm. How to practically perform such dual norm computation is therefore a natural question. We address it in this section providing a centralized algorithm to compute the TV dual norm. Beforehand, a few technical results are presented. All proofs of this section are provided in Appendix III.

A. Further Properties of $(\mathbb{R}_0^V, \|\cdot\|_*)$

Another characterization of the dual norm is the following. We are going to need an arbitrary orientation to each edge: $\vec{E}$ denotes whatever compatible set of directed edges, in the sense that $(v, w) \in \vec{E}$ implies that $\{v, w\} \in E$ and $(w, v) \notin \vec{E}$. Reciprocally $\{v, w\} \in E$ implies either $(v, w) \in \vec{E}$ or $(w, v) \in \vec{E}$. An element $\boldsymbol{\xi} \in \mathbb{R}^{\vec{E}}$ is called a *vector field*. We denote by $\|\boldsymbol{\xi}\|_\infty = \max\{|\boldsymbol{\xi}(e)| : e \in \vec{E}\}$. For a vector field $\boldsymbol{\xi} \in \mathbb{R}^{\vec{E}}$, $\text{div}\,\boldsymbol{\xi}$ denotes the function:

$$\text{div}\,\boldsymbol{\xi} : v \in V \mapsto \sum_{(v,w)\in\vec{E}} \boldsymbol{\xi}(v,w) - \sum_{(w,v)\in\vec{E}} \boldsymbol{\xi}(w,v)$$

The following proposition provides a characterization of the dual norm. Its proof is adapted from [36].

Proposition 4: If G is a connected graph, the following equality holds true:

$$\forall \boldsymbol{u} \in \mathbb{R}_0^V, \quad \|\boldsymbol{u}\|_* = \inf\{\|\boldsymbol{\xi}\|_\infty : \boldsymbol{u} = \text{div}\,\boldsymbol{\xi}\}. \qquad (14)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2015.2471755, IEEE Transactions on Automatic Control

B. Co-area Formula

The following lemma, also known in the context of real analysis as the *coarea formula*, will be helpful to prove Proposition 5.

Lemma 1: For a function $\boldsymbol{x} \in \mathbb{R}^V$, we denote by $\{\boldsymbol{x} \geq \lambda\} = \{v \in V : \boldsymbol{x}(v) \geq \lambda\}$ the upper-level set associated with level λ. The following equality holds true:

$$\|\boldsymbol{x}\|_{\mathrm{TV}} = \int_{-\infty}^{+\infty} \mathrm{Per}(\{\boldsymbol{x} \geq \lambda\})\mathrm{d}\lambda\,.$$

The following useful result can be seen as an extension of the immediate formula $\|\boldsymbol{u}\|_* = \max_{\boldsymbol{x}\in\mathbb{R}^V} \langle \boldsymbol{u}, \boldsymbol{x}\rangle / \|\boldsymbol{x}\|_{\mathrm{TV}}$. This results shows that instead of taking the max over the whole unit TV ball, one can consider a finite number of specific extremal points. Namely the indicator function of small subsets S of V properly renormalized. By small it is meant $|S| \leq |V|/2$.

Proposition 5: Assume $\boldsymbol{u}$ is in $(\mathbb{R}_0^V, \|\cdot\|_*)$. Then the following equalities hold true:

$$\begin{aligned} \|\boldsymbol{u}\|_* &= \max_{\emptyset\subsetneq S\subsetneq V} \frac{\langle \boldsymbol{u}, \mathbf{1}_S\rangle}{\|\mathbf{1}_S\|_{\mathrm{TV}}} \\ &= \max_{\substack{\emptyset\subsetneq S\subset V, |S|\leq |V|/2 \\ G(S) \text{ is connected}}} \frac{|\langle \boldsymbol{u}, \mathbf{1}_S\rangle|}{\|\mathbf{1}_S\|_{\mathrm{TV}}}\,. \end{aligned}$$

C. Dual norm computation

We now propose a strongly polynomial-time combinatorial algorithm to compute the dual norm of a vector. From Proposition 5, we know that $\|\boldsymbol{u}\|_*$ can be computed by enumeration of all subsets A of size at most $|V|/2$ inducing a connected subgraph. The number of such subsets is polynomially bounded for some classes of graphs (e.g., paths and cycles). However, in the general case, their number might not be polynomial. Another way to compute $\|\boldsymbol{u}\|_*$ consists in using either (11) or (14). Observe that (11) or (14) are linear programs that can be solved in polynomial time using any standard linear programming algorithm. In fact, (14) is simply the dual program of (11). Even if linear programming algorithms are very efficient, we will describe a strongly polynomial-time combinatorial algorithm to compute the dual norm of a vector, that is both practical and simple.

Algorithm 3 Dual Norm Computation Algorithm

Dual Norm Computation
- Input: $\boldsymbol{u} \in \mathbb{R}_0^V$
- Initialize $A_0 \subset V$, $\lambda_0 \leftarrow$ None
- **repeat**
  - Update $i \leftarrow i+1$
  - Update $\lambda_i = \frac{\langle \boldsymbol{u}, \mathbf{1}_{A_{i-1}}\rangle}{\mathrm{Per}(A_{i-1})}$
  - Update $A_i = \arg\max_{A\subset V} \langle \boldsymbol{u}, \mathbf{1}_A\rangle - \lambda_i \,\mathrm{Per}(A)$
- **until** $\lambda_i = \lambda_{i-1}$

Details related to the computation of $\arg\max_{A\subset V} \langle \boldsymbol{u}, \mathbf{1}_A\rangle - \lambda_i \,\mathrm{Per}(A)$ will be given just after the following proposition, proved in Appendix III-E.

Proposition 6: Algorithm 3 stops after at most $O(|E|)$ iterations. The sought output, $\|\boldsymbol{u}\|_*$ is given by the value of λ_i at the last iteration.

In order to make **Algorithm 3** practical, we still must specify how to solve the subproblem $\max_{A\subset V} \langle \boldsymbol{u}, \mathbf{1}_A\rangle - \lambda \,\mathrm{Per}(A)$. Let us now mention how this subproblem reduces to a standard *max-flow/min-cut* problem [37].

D. Reduction to max-flow/min-cut

Recall that a network (N, K, c) in graph theory sense is defined by a directed graph (N, K) and a capacity assignment $c_{(v,w)} \geq 0$ for any link $(v, w) \in K$.

Given the undirected graph $G = (V, E)$ and a vector $\boldsymbol{u} \in \mathbb{R}_0^V$, we build a network $(N = V \cup \{s, t\}, K, c)$ as follows. For each edge $\{w, v\} \in E$ we create two directed edges (w, v) and (v, w) each of capacity $c_{(v,w)} = c_{(w,v)} = \lambda$. In addition to all nodes of V, we add two other nodes: a source s and a sink t. Given any node v, if $\boldsymbol{u}(v) > 0$, we create a directed edge (s, v) of capacity $\boldsymbol{u}(v)$, while an arc (v, t) of capacity $|\boldsymbol{u}(v)|$ is added if $\boldsymbol{u}(v) < 0$. Thus, the set of links of the network is given by $K = \{(v, w), \{v, w\} \in E\} \cup \{(s, v), \boldsymbol{u}(v) > 0\} \cup \{(v, t), \boldsymbol{u}(v) < 0\}$.

Let A be any subset of vertices of V and let $\delta^+(A \cup \{s\})$ denote the set of directed edges having only their first extremity in $A \cup \{s\}$. Quantity $\delta^+(A \cup \{s\})$ is generally called a cut. This cut separates s and t in the sense that $s \in A \cup \{s\}$ while $t \notin A \cup \{s\}$.

The capacity of this cut is defined as the sum of the capacities of the directed edges included in the cut. Let us denote it by $c(\delta^+(A \cup \{s\}))$. It is easy to see that $c(\delta^+(A \cup \{s\}))$ is equal to:

$$\begin{aligned} & \lambda\,\mathrm{Per}(A) - \sum_{v\in A, \boldsymbol{u}(v)<0} \boldsymbol{u}(v) + \sum_{v\in V\setminus A, \boldsymbol{u}(v)>0} \boldsymbol{u}(v) \\ = \;& \lambda\,\mathrm{Per}(A) - \sum_{v\in A} \boldsymbol{u}(v) + \sum_{v\in A, \boldsymbol{u}(v)>0} \boldsymbol{u}(v) + \sum_{v\in V\setminus A, \boldsymbol{u}(v)>0} \boldsymbol{u}(v) \\ = \;& \lambda\,\mathrm{Per}(A) - \langle \boldsymbol{u}, \mathbf{1}_A\rangle + \sum_{v\in V, \boldsymbol{u}(v)>0} \boldsymbol{u}(v) \end{aligned}$$

Observe that $c(\delta^+(A \cup \{s\}))$ is the sum of the term $\lambda\,\mathrm{Per}(A) - \langle \boldsymbol{u}, \mathbf{1}_A\rangle$ and a constant term not depending on A. Then, computing a minimum-capacity cut is clearly equivalent to finding a subset A maximizing $\max_{A\subset V} \langle \boldsymbol{u}, \mathbf{1}_A\rangle - \lambda\,\mathrm{Per}(A)$. A minimum-capacity cut can be computed using any maximum-flow/minimum-cut algorithm such as the Edmonds-Karp's algorithm, the Goldberg-Tarjan's algorithm or Orlin's Algorithm (see, e.g., [37]).

Since each iteration of Algorithm 3 calls such a maximum-flow subroutine, and Proposition 6 asserts that there is at most $|E|$ iterations (see the appendix for a proof), the overall complexity of Algorithm 3 is consequently given by $|E|$ times the complexity of the maximum-flow algorithm (which depends on the algorithm used).

VI. Stubborn Agents

One of the claims of this paper is that the above algorithms are attractive in order to provide robustness against misbehaving agents. This claim is motivated by the example below. Assume that some agents, called *stubborn*, never change their state. The rationale behind this model is twofold: either these agents are malfunctioning, or they might want to deliberately pollute or influence the network.

For ease of interpretation, we will focus on the average consensus case. We represent the state vector as $\boldsymbol{x}_n = (\boldsymbol{x}_n^R, \boldsymbol{x}_n^S)$ where $\boldsymbol{x}_n^R$ (resp. $\boldsymbol{x}_n^S$) is the restriction of $\boldsymbol{x}_n$ to regular agents (resp. stubborn agents). By definition, $\boldsymbol{x}_n^S = \boldsymbol{x}_0^S$ for any n.

A. Failure of Linear Gossip

Assume that the state vector $\boldsymbol{x}_n$ is written in block form $\boldsymbol{x}_n = \begin{pmatrix} \boldsymbol{x}_n^R \\ \boldsymbol{x}^S \end{pmatrix}$. The state vector is updated according to standard linear gossip scheme $\boldsymbol{x}_{n+1} = \boldsymbol{W}\boldsymbol{x}_n$ where $\boldsymbol{W}$ is a square matrix. We make the following natural assumptions about the matrix $\boldsymbol{W}$.

Assumption 3 (Linear Gossip Structure):

(a) Matrix $\boldsymbol{W}$ has the following structure;

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{W}^R & \boldsymbol{W}^S \\ 0 & \boldsymbol{I} \end{pmatrix} \tag{15}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2015.2471755, IEEE Transactions on Automatic Control

(b) Matrix $\boldsymbol{W}$ is a right stochastic matrix: its entries are in $[0,1]$ and $\boldsymbol{W}\mathbf{1}_V = \mathbf{1}_V$.
(c) Considering the directed edge structure E' defined by: $(v,w) \in E' \Leftrightarrow \boldsymbol{W}(v,w) > 0$, there exists a directed path from each regular node r to at least one stubborn node s.

The block structure of matrix W follows from the constraint that *stubborn* agents do not change their state over time. The last requirement is also very natural. If it were to be not fulfilled, there would exist regular agents that communicate in autarky and cannot be aware of *stubborn* agents' opinions.

We now address the issue of convergence of such gossip algorithms in the presence of *stubborn* agents. The following result is proved in [38].

Proposition 7 ([38]): Under Assumption 3, algorithm $\boldsymbol{x}_{n+1} = \boldsymbol{W}\boldsymbol{x}_n$ converges to

$$\boldsymbol{x}_\infty = (\boldsymbol{I} - \boldsymbol{W}^R)^{-1}\boldsymbol{W}^S\boldsymbol{x}^S \tag{16}$$

Consensus is not necessarily reached since $\boldsymbol{x}_\infty$ is not proportional to $\mathbf{1}_V$ as long as *stubborn* agents disagree with each other. In addition, the limit state does not depend on the initial state, it only depends on the *stubborn* agents state. In other terms, the *stubborn* agents solely drive the network, initial opinions of regular agents is lost, even with one single *stubborn* agent.

B. ADMM and subgradient algorithm

As a straightforward extension of Theorems 2 and 3, it is not difficult to see that in the presence of stubborn agents, the sequence $\boldsymbol{x}_n^R$ generated either by Algorithms IV-A or IV-B converges to the minimizers of the following *perturbed* optimization problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^R} \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{x}_0^R\|_2^2 + \lambda\,\|\boldsymbol{x}\|_{\mathrm{TV}} + \lambda \sum_{\substack{v\in R, w\in S\\ v\sim w}} |\boldsymbol{x}(v) - \boldsymbol{x}_0(w)|\,. \tag{17}$$

where, here, $\|\boldsymbol{x}\|_{\mathrm{TV}}$ is to be understood as the total variation of a function $\boldsymbol{x} \in \mathbb{R}^R$ on the subgraph $G(R)$ *i.e.*, the restriction of G to the set of regular agents. To ease the reading and with no risk of ambiguity, we still keep the same notations $\|\,.\,\|_{\mathrm{TV}}$ and $\|\,.\,\|_*$ to designate the TV and the dual norms associated with $G(R)$.

A complete proof of robustness of our algorithm would require a closed-form expression of the minimizers of (17). Unfortunately, such an explicit characterization is a difficult task. In the general case, solving analytically problem (17) seems unfortunately out of reach. Therefore, in order to validate the claim that our algorithms are indeed robust, we are left with two options. First, we should provide extensive numerical results that exhibits the robustness in practical scenarios. This is done in Section VII. Second, we must prove robustness *in some case study* for which the minimizers of (17) are tractable. In the sequel, we characterize the minimizers in the following simplified scenario.

Scenario 1: Any stubborn agent is connected to all regular agents. In addition, there exists $a \in \mathbb{R}$ such that $\boldsymbol{x}_0(s) = a$ for any $s \in S$.
Loosely speaking, one might think of Scenario 1 as a worst-case situation in the sense that each stubborn agent directly disturbs *all* regular agents.

Let $\bar{\boldsymbol{x}}_0^R = \frac{1}{|R|}\sum_{v\in R}\boldsymbol{x}_0^R(v)$. The following result is proved in Appendix IV-A.

Theorem 4: Assume that $\lambda \geq \|\boldsymbol{x}_0^R - \bar{\boldsymbol{x}}_0^R 1_R\|_*$ and let

$$x^* = \begin{cases} a & \text{if } |\bar{\boldsymbol{x}}_0^R - a| \leq \lambda|S| \\ \bar{\boldsymbol{x}}_0^R + \lambda|S| & \text{if } \bar{\boldsymbol{x}}_0^R + \lambda|S| < a \\ \bar{\boldsymbol{x}}_0^R - \lambda|S| & \text{if } \bar{\boldsymbol{x}}_0^R - \lambda|S| > a \end{cases}.$$

Then, in Scenario 1, $x^*\mathbf{1}_R$ is the unique minimizer of Problem (17).

Observe that even if the common value a of the *stubborn* coalition is very far from $\bar{\boldsymbol{x}}_0^R$, we will reach a consensus within a distance $\lambda|S|$ from $\bar{\boldsymbol{x}}_0^R$. The same conclusion holds when a is already close to $\bar{\boldsymbol{x}}_0^R$ (we still reach a consensus within a distance $\lambda|S|$ from $\bar{\boldsymbol{x}}_0^R$). The quantity $\lambda|S|$ can be interpreted as the robustness level of our algorithms in the sense of a maximum error margin. Therefore the proposed algorithm is unlike more standard gossip algorithms which can be driven arbitrarily far away from the sought consensus. Note that a small λ reduces the error margin. The tradeoff is related to the fact that the selection of a small λ also reduces the set of functions $\boldsymbol{x}_0^R$ satisfying the regularity condition $\|\boldsymbol{x}_0^R - \bar{\boldsymbol{x}}_0^R\mathbf{1}_R\|_* \leq \lambda$.

For theoretical results concerning other scenarios, one should carefully rework expression 17. On a more intuitive level, remark that this expression is the sum of two terms: $\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{x}_0^R\|_2^2 + \lambda\,\|\boldsymbol{x}\|_{\mathrm{TV}}$ and $\lambda\sum_{\substack{v\in R, w\in S\\ v\sim w}}|\boldsymbol{x}(v)-\boldsymbol{x}_0(w)|$. Only the second term depends on the stubborn agents. If one imagines that the regular agents values remain fixed and that the stubborn agents start taking large values, the overall value varies like $O(\lambda)$ (the constant depending on the number of stubborn agents involved); while if one changes the regular values such that $\bar{\boldsymbol{x}}^R$ changes then a quadratic increase is paid in the first term. This hints at why large variations of isolated stubborn agents should not shift regular agents average value too much.

VII. Numerical Experiments

In this section, the previous results are illustrated numerically. We first validate that, in a stubborn-free network, both Algorithm IV-A and Algorithm IV-B do converge to some x^*, minimizer of (8). Since eq. (8) involves abstract functions f_v, we consider two choices, namely $f_v(x) = (x - \boldsymbol{x}_0(v))^2$ and $f_v(x) = |x - \boldsymbol{x}_0(v)|$. Notice that in the average consensus case there is a single minimizer since the energy is strictly convex. We are able to predict the minimizers of (8) by computing $\|\partial F(x^*\mathbf{1}_V)\|_*$, for $x^* = \bar{\boldsymbol{x}}_0$ in the case where $f_v(x) = (x - \boldsymbol{x}_0(v))^2$, and $x^* \in \operatorname{median}\boldsymbol{x}_0$ in the case where $f_v(x) = |x - \boldsymbol{x}_0(v)|$, using Algorithm 3 and then invoking Theorem 1. We then introduce stubborn agents and validate Theorem 4 numerically checking that whenever $\lambda \geq \|\boldsymbol{x}_0^R - \bar{\boldsymbol{x}}_0^R\mathbf{1}_R\|_*$, regular agents do achieve consensus, settling the case of perturbed average consensus.

A. Framework

Unless stated otherwise, the underlying network is the complete graph with $N = 99$ agents (to have a unique median). In order to represent the data over the network, such as $v \in V \mapsto \boldsymbol{x}(v)$, we choose an arbitrary order among the vertices, so that we can identify set V with $\{1,\ldots,N\}$. We then plot function $v \in \{1,\ldots,N\} \mapsto \boldsymbol{x}(v)$. Again for the sake of simplicity, whenever set S is not empty – *i.e.* when there are some stubborn agents in the network – we always assume there is a single stubborn agent indexed by 1, namely: $S = \{1\}$.

For the sake of simplicity we do not vary the parameters of Algorithms IV-A and IV-B. In all experiments, we used $\gamma_n = (n+1)^{-1}$ for the subgradient descent and $\rho = 1$ for the ADMM.

The initial data is represented by circles in Figure 1. The average value of regular agents is approximately $\bar{\boldsymbol{x}}_0 = 0.1271$ and 0.2817 is a median. In the case where there is a stubborn agent, the initial data is the same except for the first agent that is supposed stubborn. Three possible values will be considered for the stubborn agent ($\boldsymbol{x}_0(1) = 10$, -10 or 0.16).

B. Convergence to the minimizers of eq. (8)

We consider two cases: the average consensus where $f_v(x) = (x - \boldsymbol{x}_0(v))^2$ and the median consensus problem $f_v(x) = |x - \boldsymbol{x}_0(v)|$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2015.2471755, IEEE Transactions on Automatic Control

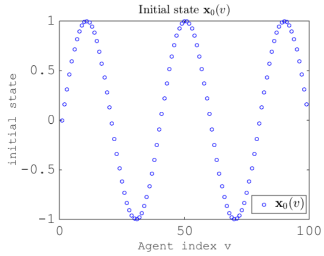

Fig. 1. Initial data (circles) as a function of agent index.

In the average consensus case, the energy to minimize is strictly convex and admits a unique minimizer. In the median consensus case, convexity is not strict, and uniqueness does not necessarily hold.

For both cases, we provide two plots. The first plot shows function $n \mapsto \log \|J^{\perp} \boldsymbol{x}_n\|$, where operator $J^{\perp} = I_N - J$, with I_N the $n \times n$ identity matrix and J the orthogonal projector on the consensus subspace (hence $J^{\perp}$ is the orthogonal projector on $1^{\perp}$). In other terms $\|J^{\perp} \boldsymbol{x}_n\|$ measures how much disagreement there is left in the network. Please note the log in front, meaning that if such a function decrease linearly, it implies that $\boldsymbol{x}_n$ goes to consensus exponentially fast. The second plot shows functions $n \mapsto \bar{\boldsymbol{x}}_n$ – except in the average consensus case – since the graph is $N-1$-regular, $\bar{\boldsymbol{x}}_n$ is kept constant along the iterations when using both Algorithms IV-A and IV-B.

1) Average Consensus: In the regular case, we compute the target $\bar{\boldsymbol{x}}_0 = 0.1271$. Next, we compute $\|\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0 \mathbf{1}_V\|_* = 0.0130$ using Algorithm 3. We set $\lambda = 0.05 > \|\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0 \mathbf{1}_V\|_*$. Figure 2 (left) shows function $n \mapsto \log \|J^{\perp} \boldsymbol{x}_n\|$ for both Algorithms IV-A and IV-B. Our results predict and the simulations confirm that $n \mapsto \bar{\boldsymbol{x}}_n$ remain constant with n. It appears that Algorithm IV-B goes to consensus exponentially fast.

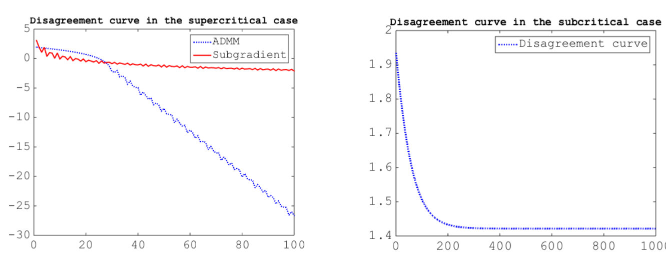

Fig. 2. Average consensus. Left: disagreement measure $n \to \log \|J^{\perp} \boldsymbol{x}_n\|$ for Algorithms IV-A and IV-B. Function $n \to \bar{\boldsymbol{x}}_n$ is not represented since it is kept constant along the iterations. Note how faster Algorithm IV-B converges to consensus. Right: Case where $\lambda < \|\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0 \mathbf{1}_V\|_*$; consensus is not achieved and the quantity of disagreement, measured by $\log \|J^{\perp} \boldsymbol{x}_n\|$ stabilizes after approximately 200 iterations.

On the opposite, we compare with the situation where λ is subcritical, *i.e.* less than $\|\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0 \mathbf{1}_V\|_* = 0.0130$. We set $\lambda = 0.005$. Figure 2 (right) represents the curve $n \to \log \|J^{\perp} \boldsymbol{x}_n\|$ for Algorithm IV-B which appears the fastest. One can observe that consensus is not achieved, since $\log \|J^{\perp} \boldsymbol{x}_n\|$ stays approximately constant after 200 iterations instead of going down in a linear fashion as in the supercritical case, *i.e.* when $\lambda > \|\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0 \mathbf{1}_V\|_*$. The subcritical scenario is of interest when stubborn agents are present or more generally where the agents underestimated their initial disagreement $\|\boldsymbol{x}_0 - \bar{\boldsymbol{x}}_0 \mathbf{1}\|_*$ which results in the failure of reaching an asymptotic consensus.

In the case of supercritical $\lambda > \lambda_*$, it is also of interest to change the graph topology and study its impact on convergence to consensus. We focus on two important classes of random graphs: Erdös-Rényi and Random Geometric Graphs. Erdös-Rényi graphs are denoted $G(n, p)$, n being the number of vertices and p being the probability of edges (all independent with the same Bernoulli distribution with parameter p). Random Geometric graphs are denoted $G(d, r, n)$ where d denotes the dimension of the underlying cube, r the critical radius and n the number of vertices. For both families it is known that when p (resp. r) is sufficiently large the random graph is connected with high probability [39]. In figure 3, one can notice the impact of connectivity: when p (resp. r) increases, convergence to consensus seems to takes place more rapidly.

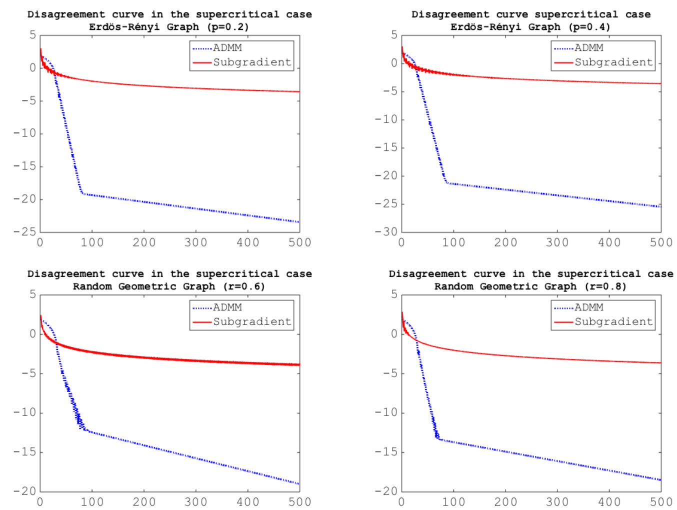

Fig. 3. Average consensus. Disagreement measure $n \to \log \|J^{\perp} \boldsymbol{x}_n\|$ for Algorithms IV-A and IV-B. Top: Erdös-Rényi graph. Bottom: Random Geometric graphs. Left: Less connected graphs. Right: More connected graphs. Precisely from left to right and top to bottom $G(n, p_1)$, $G(n, p_2)$, $RGG(2, r_1, n)$ and $RGG(2, r_2, n)$ with $n = 99$, $p_1 = .2$, $p_2 = .4$, $r_1 = .6$ and $r_2 = .8$.

2) Median Consensus: In the regular case, we compute the target median $\boldsymbol{x}_0 = 0.2817$. For any non-empty subset $A \subset V$ in the complete graph it is easy to see that $\mathrm{Per}(A) \geq N - 1$. Hence, using Proposition 5 it is easy to see that $\lambda_0 \leq N/(2N - 2)$. In order to satisfy $\lambda > \lambda_0$ we choose $\lambda = .5$.

Figure 4 shows the two curves $n \mapsto \log \|J^{\perp} \boldsymbol{x}_n\|$ and $n \mapsto \bar{\boldsymbol{x}}_n$ for both Algorithms IV-A and IV-B. One can notice that Algorithm IV-A has not converged after 350 iterations while Algorithm IV-B has. One can also notice, comparing the slope breaks between the left and right figures that Algorithm IV-B goes to consensus fast and then evolves at consensus towards the target value.

For the sake of conciseness, we do not vary the graph structure in the median case. However the qualitative results are the same than in the average case: namely, the more connected the graph is, the faster consensus seems to be reached.

C. Behavior in the presence of stubborn agents

We now investigate numerically the case when stubborn agents are introduced in the network, and more specifically the case where one single stubborn agent is introduced and the graph is the complete graph to validate Theorem 4. We use the same initialization state as in the previous experiments, except that the first agent is assumed stubborn. The target value in this case is $\bar{\boldsymbol{x}}_0^R = 0.1284$. Then we distinguish three cases: $\boldsymbol{x}_0(1) = 10$, $\boldsymbol{x}_0(1) = 0.16$ and $\boldsymbol{x}_0(1) = -10$. When $\lambda = 0.05 \geq \|\boldsymbol{x}_0^R - \bar{\boldsymbol{x}}_0^R\|_* = 0.0133$ (as computed by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2015.2471755, IEEE Transactions on Automatic Control

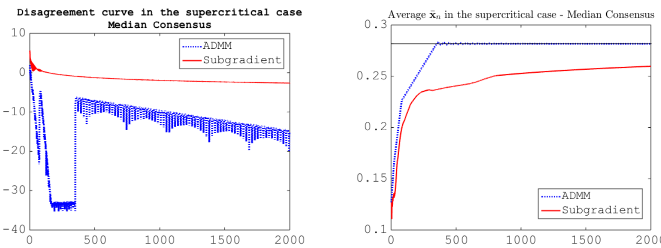

Fig. 4. Median consensus. Left: disagreement measure $n \to \log \|J^{\perp}\boldsymbol{x}_n\|$ for Algorithms IV-A and IV-B. Right: function $n \to \bar{\boldsymbol{x}}_n$ for Algorithm IV-A and IV-B, and a horizontal line corresponding to $\mathrm{median}(\boldsymbol{x}_0)$. Observe an interesting feature of Algorithm IV-B: it reaches the target value at about iteration 350 and then starts slightly oscillating around it; while Algorithm IV-A does not converge after 2 000 iterations.

Algorithm 3), there should be convergence to consensus to $\bar{\boldsymbol{x}}_0^R + \lambda$ in the first case, $\boldsymbol{x}_0(1)$ in the second case and $\bar{\boldsymbol{x}}_0^R - \lambda$ in the third case; for both Algorithms. This is exactly what is found in the numerical experiments we performed as reported in Figure 5; only experiments using Algorithm IV-B are reported to avoid too much clutter.

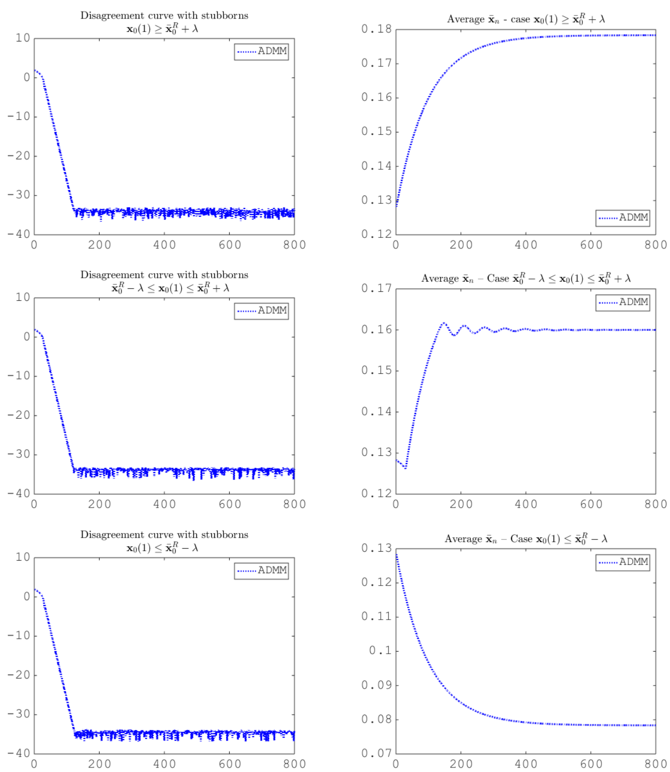

Fig. 5. With a stubborn agent. Left: disagreement measure $n \to \log \|J^{\perp}\boldsymbol{x}_n\|$ for Algorithm IV-B. Right: function $n \to \bar{\boldsymbol{x}}_n$ for Algorithm IV-B. First line stubborn agent has value $\boldsymbol{x}_0(1) = 10$; $\bar{\boldsymbol{x}}_n$ converges to .1784 which corresponds to $\bar{\boldsymbol{x}}_0^R + \lambda$ up to precision 10^{-4}. Mid line stubborn has value $\boldsymbol{x}_0(1) = 0.16$; again, well aligned with Theorem 4 since $\bar{\boldsymbol{x}}_n$ converges to 0.16. Third line, $\boldsymbol{x}_0(1) = -10$; $\bar{\boldsymbol{x}}_n$ converges to 0.0784 that corresponds to $\bar{\boldsymbol{x}}_0^R - \lambda$.

We also consider more stubborn agents and other graph topology in figure 6. More specifically we consider respectively Erdös-Rényi and Random Geometric graphs with 10% stubborn agents with common value that is large compared the average of the network. Notice how the consensus remains of order 0.1 in each case.

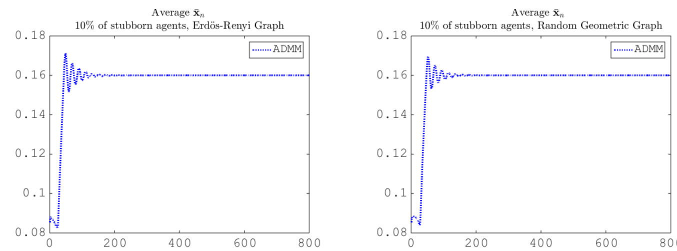

Fig. 6. With 10 stubborn agent and other graph topologies. Function $n \to \bar{\boldsymbol{x}}_n$ for Algorithm IV-B is plotted. Stubborn agents common value is still 10. Left: Erdös-Rényi graph $G(99, .1)$. Right: Random Geometric Graph $G(2, .6, 99)$.

VIII. Conclusion

In this paper we analyzed the so called consensus distributed optimization problem, that writes $\inf_{\boldsymbol{x}} \sum_{v \in V} f_v(x) + \iota_{C}(\boldsymbol{x})$. We have shown that it could be beneficial to consider instead $\inf_{\boldsymbol{x}} \sum_v f_v(\boldsymbol{x}(v)) + \lambda \|\boldsymbol{x}\|_{\mathrm{TV}}$ that corresponds to a relaxation of the consensus distributed optimization problem. A relaxation in the sense that consensus is no more imposed as a hard constraint, just encouraged using the TV term. Firstly, we have shown that, remarkably enough, a critical threshold λ^* involving the dual of the TV norm could be computed. When λ is set above λ^*, both problems actually yield to the same minimizers. Secondly, we proposed two distributed algorithms to reach these minimizers. Thirdly we proposed a centralized algorithm to compute the critical λ^*. And finally, we established another noticeable feature of the TV regularization framework: it is more robust to stubborn agents in the network. We demonstrated this in a formal way using a specific assumption on the network and also using numerical experiments. This work advocates the use of TV regularization as a tool for performing distributed optimization. It opens interesting perspectives and raises several questions: is it possible to match convergence rates of consensus distributed optimization algorithms? is it possible to prove more general robustness properties of TV regularization with respect to stubborn agents or other forms of misbehaving agents?

References

[1] Leonid I Rudin, Stanley Osher, and Emad Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1, pp. 259–268, 1992.

[2] Emmanuel J Candes and Terence Tao, "Decoding by linear programming," *Information Theory, IEEE Transactions on*, vol. 51, no. 12, pp. 4203–4215, 2005.

[3] Emmanuel J Candès, Justin Romberg, and Terence Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, no. 2, pp. 489–509, 2006.

[4] David L Donoho, "Compressed sensing," *IEEE Transactions on Information Theory*, vol. 52, no. 4, pp. 1289–1306, 2006.

[5] Luigi Ambrosio, Nicola Fusco, and Diego Pallara, *Functions of bounded variation and free discontinuity problems*, Oxford Press, 2000.

[6] Abderrahim Elmoataz, Olivier Lezoray, and Sébastien Bougleux, "Nonlocal discrete regularization on weighted graphs: a framework for image and manifold processing," *Image Processing, IEEE Transactions on*, vol. 17, no. 7, pp. 1047–1060, 2008.

[7] Camille Couprie, Leo Grady, Laurent Najman, Jean-Christophe Pesquet, and Hugues Talbot, "Dual constrained tv-based regularization on graphs," *SIAM Journal on Imaging Science*, vol. 6, no. 3, pp. 1246–1273, 2013.

[8] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, 1997.

[9] C. Lopes and A.H. Sayed, "Distributed processing over adaptive networks," in *Adaptive Sensor Array Processing Workshop*, June 2006, pp. 1–5.

[10] S.S. Ram, A. Nedic, and V. Veeravalli, "Distributed stochastic subgradient projection algorithms for convex optimization," *Journal of Optimization Theory and Applications*, vol. 147, pp. 516–545, 2010.

[11] P. Bianchi and J. Jakubowicz, "Convergence of a multi-agent projected stochastic gradient algorithm for non-convex optimization," *IEEE Trans. on Automatic Control*, vol. 58, no. 2, 2013.

[12] D. Kempe, A. Dobra, and J. Gehrke, "Gossip-based computation of aggregate information," in *Foundations of Computer Science, Annual IEEE Symposium on*. 2003, IEEE Computer Society.

[13] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Randomized Gossip Algorithms," *IEEE Transactions on Inform. Theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[14] Dusan Jakovetic, Joao Xavier, and Jose MF Moura, "Fast distributed gradient methods," *arXiv preprint arXiv:1112.2972*, 2011.

[15] John C Duchi, Alekh Agarwal, and Martin J Wainwright, "Dual averaging for distributed optimization: convergence analysis and network scaling," *IEEE Transactions on Automatic Control*, vol. 57, no. 3, pp. 592–606, 2012.

[16] I.D. Schizas, A. Ribeiro, and G.B. Giannakis, "Consensus in ad hoc WSNs with noisy links-Part I: Distributed estimation of deterministic signals," *IEEE Transactions on Signal Processing*, vol. 56, no. 1, pp. 350–364, 2008.

[17] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Now Publishers Inc., 2011.

[18] F. Iutzeler, P. Bianchi, Ph. Ciblat, and W. Hachem, "Asynchronous distributed optimization using a randomized alternating direction method of multipliers," in *IEEE Conf. on Decision and Control*, 2013.

[19] C. Eksin and A. Ribeiro, "Distributed network optimization with heuristic rational agents," *IEEE Trans. Signal Process,(submitted)*, 2011.

[20] D. Acemoglu, G. Como, F. Fagnani, and A. Ozdaglar, "Opinion fluctuations and persistent disagreement in social networks," in *IEEE CDC 2011*, 2011, pp. 2347–2352.

[21] A-R Mawlood-Yunis, Michael Weiss, and Nicola Santoro, "Issues for robust consensus building in p2p networks," in *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*. Springer, 2006, pp. 1021–1027.

[22] W. Ben-ameur, P. Bianchi, and J. Jakubowicz, "Robust average consensus using total variation gossip algorithm," in *ValueTools' 12: 6th International Conference on Performance Evaluation Methodologies and Tools*, 2012.

[23] Fabio Pasqualetti, Antonio Bicchi, and Francesco Bullo, "Consensus computation in unreliable networks: A system theoretic approach," *IEEE Transactions on Automatic Control*, vol. 57, no. 1, pp. 90–104, 2012.

[24] Meng Guo, Dimos V Dimarogonas, and Karl Henrik Johansson, "Distributed real-time fault detection and isolation for cooperative multi-agent systems," in *American Control Conference (ACC), 2012*. IEEE, 2012, pp. 5270–5275.

[25] Peng Lin, Yingmin Jia, and Lin Li, "Distributed robust hinfinity consensus control in directed networks of agents with time-delay," *Systems & Control Letters*, vol. 57, no. 8, pp. 643–653, 2008.

[26] Haotian Zhang and Shreyas Sundaram, "Robustness of information diffusion algorithms to locally bounded adversaries," in *American Control Conference (ACC), 2012*. IEEE, 2012, pp. 5855–5861.

[27] F. Lin, M. Fardad, and M.R. Jovanovic, "Design of optimal sparse feedback gains via the alternating direction method of multipliers," *Automatic Control, IEEE Transactions on*, vol. 58, no. 9, pp. 2426–2431, Sept 2013.

[28] J.-B. Hiriart-Urruty and C. Lemaréchal, *Convex Analysis and Minimization Algorithms. Part 1: Fundamentals*, vol. 305, Springer, 1996.

[29] F. R. K. Chung, *Spectral Graph Theory*, American Mathematical Society, 1997.

[30] Francis Bach, Rodolphe Jenatton, Julien Mairal, and Guillaume Obozinski, "Optimization with sparsity-inducing penalties," *arXiv preprint arXiv:1108.0775*, 2011.

[31] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," Tech. Rep., October 2003, Lecture Notes.

[32] Naum Zuselevich Shor, *Minimization methods for non-differentiable functions*, Springer-Verlag, 1979.

[33] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87, Springer, 2004.

[34] Patrick L. Combettes and Jean christophe Pesquet, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, chapter Proximal Splitting Methods in Signal Processing, pp. 185–222, Springer, 2011.

[35] Tomaso Erseghe, Davide Zennaro, Emiliano Dall'Anese, and Lorenzo Vangelista, "Fast consensus by the alternating direction multipliers method," *Signal Processing, IEEE Transactions on*, vol. 59, no. 11, pp. 5523–5537, 2011.

[36] R. A. Adams, *Sobolev spaces*, Academic Press, 1975.

[37] B. H. Korte and J. Vygen, *Combinatorial optimization*, vol. 21, Springer, 2012.

[38] U.A Khan, S. Kar, and J.M.F. Moura, "Higher dimensional consensus: Learning in large-scale networks," *IEEE Transactions on Signal Processing*, vol. 58, no. 5, pp. 2836–2849, May 2010.

[39] B. Bollobás, *Random graphs*, Springer, 1998.

Some results are used in the appendices and proved afterwards for the sake of readability.

Appendix I
Proofs of Section III

A. Proof of Theorem 1

Proof: [1) $\Leftrightarrow$ 2)] Note that $x^\star \mathbf{1}_V$ is a minimizer of $F + \lambda\|\cdot\|_{\mathrm{TV}}$ iff $\mathbf{0} \in \partial F(x^\star \mathbf{1}_V) + \lambda \partial \|x^\star \mathbf{1}_V\|_{\mathrm{TV}}$. From Proposition 4, $\partial\|x^\star \mathbf{1}_V\|_{\mathrm{TV}} = B_*$. Therefore, 1) holds iff there exists $\boldsymbol{u} \in \partial F(x^\star \mathbf{1}_V)$ such that $\mathbf{0} \in \boldsymbol{u} + \lambda B_*$. Otherwise stated, there exists $\boldsymbol{u} \in \partial F(x^\star \mathbf{1}_V)$ such that $\boldsymbol{u} \in \lambda B_*$.

[2) $\Leftrightarrow$ 3)] is a consequence of Proposition 5.

[3) $\Rightarrow$ 4)] As $\sum_v \partial f_v = \partial(\sum_v f_v)$, condition $\sum_{v\in V} \boldsymbol{u}(v) = 0$ implies that $0 \in \partial(\sum_v f_v)(x^\star)$. Thus, $x^\star$ is a minimizer of $\sum_v f_v$. ∎

B. Proof of Proposition 2

Proof: Uniqueness of the minimizers is a consequence of the strict convexity, while the equivalence of the three statements follows directly from Theorem 1. ∎

C. Proof of Proposition 3

Proof: Consider any $x^\star \in \mathrm{median}(\boldsymbol{x}_0)$. Consider a bijection $\sigma : \{1 \cdots |V|\} \to V$ such that $(\boldsymbol{x}_0 \circ \sigma)(1) \le \cdots \le (\boldsymbol{x}_0 \circ \sigma)(|V|)$. Define $\boldsymbol{u} \in \mathbb{R}^V$ as follows. When $|V|$ is odd, $\boldsymbol{u}(v)$ is equal to 1 if $\sigma^1(v) < \frac{|V|+1}{2}$, to 0 if $\sigma^{-1}(v) = \frac{|V|+1}{2}$ and to -1 otherwise. When $|V|$ is even, $\boldsymbol{u}(v)$ is equal to 1 if $\sigma^{-1}(v) \le \frac{|V|}{2}$ and to -1 otherwise. It is straightforward to verify that $\boldsymbol{u} \in \partial F(x^\star \mathbf{1}_V)$. As $\boldsymbol{u} \in \mathcal{U}$, one has $\|\boldsymbol{u}\|_* \le \lambda$. Thus $\partial F(x^\star \mathbf{1}_V) \cap \lambda B_*$ is nonempty. By Theorem 1, $x^\star$ is a minimizer of (9).

By Theorem 1 again, all minimizers of (9) which belong to $\boldsymbol{C}$ necessarily correspond to minimizers of (1). Thus the set of minimizers of (9) which belong to $\boldsymbol{C}$ is equal to $\mathrm{median}(\boldsymbol{x}_0)\mathbf{1}_V$. It remains to show that (9) has no minimizers outside the consensus subspace $\boldsymbol{C}$. Denote by $B_0 = \{\boldsymbol{u} \in \mathbb{R}_0^V : \|\boldsymbol{u}\|_\infty \le 1\}$. From Lemma 2, its extremal points are given by set $\mathcal{U}$. Recall that, by definition,

$$\lambda_0 = \max\{\|\boldsymbol{u}\|_* : \boldsymbol{u} \in \mathcal{U}\} .$$

Since B_0 is a polytope (bounded intersection of halfspaces), it is well known that it is the convex hull of its extremal points. Triangular inequality implies in turn:

$$\lambda_0 = \max\{\|\boldsymbol{u}\|_* : \boldsymbol{u} \in B_0\} .$$

Now, assume that function $\boldsymbol{x} : V \to \mathbb{R}$ is a minimizer such that $\|\boldsymbol{x}\|_{\mathrm{TV}} > 0$. Then for some $\boldsymbol{g} \in \partial F(\boldsymbol{x})$, one has $\boldsymbol{g} = -\lambda \boldsymbol{u}$ with $\boldsymbol{u} \in \mathbb{R}_0^V$ such that $\langle \boldsymbol{u}, \boldsymbol{x} \rangle = \|\boldsymbol{x}\|_{\mathrm{TV}}$ and $\|\boldsymbol{u}\|_* \le 1$ (by Proposition 1). On the one hand, this implies that $\boldsymbol{g} \in B_0$ since $\|\partial F(\cdot)\|_\infty \le 1$ and on the other hand it implies that $\|\boldsymbol{g}\|_* = \lambda$. This contradicts $\lambda > \lambda_0$. ∎

Let us recall a standard definition and derive an easy lemma.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2015.2471755, IEEE Transactions on Automatic Control

Definition 1: Assume C is a convex set. A point $\boldsymbol{p} \in C$ is said *extremal* when:

$$\lambda \in (0,1)\,, \boldsymbol{x} \in C\,, \boldsymbol{y} \in C\,, \boldsymbol{p} = \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y} \Rightarrow \boldsymbol{x} = \boldsymbol{y}$$

Lemma 2: The polyhedral set $B_0 = \{\boldsymbol{u} \in \mathbb{R}_0^V : \|\boldsymbol{u}\|_\infty \leq 1\}$ has extremal point set $\mathcal{U}$, as defined in Section III-D.

Proof: Assume $\boldsymbol{u}$ is an extremal point of B_0. And assume by contradiction that there exists $(v, v') \in V^2$ such that: $v \neq v'$ and $\max(|\boldsymbol{u}(v)|, |\boldsymbol{u}(v')|) < 1$. Denote by $\epsilon = 1 - \max(|\boldsymbol{u}(v)|, |\boldsymbol{u}(v')|) > 0$. Then one has $\boldsymbol{u} = \frac{1}{2}(\boldsymbol{u} + \epsilon\boldsymbol{\delta}_v - \epsilon\boldsymbol{\delta}_{v'}) + \frac{1}{2}(\boldsymbol{u} - \epsilon\boldsymbol{\delta}_v + \epsilon\boldsymbol{\delta}_{v'})$, where $\boldsymbol{\delta}_v$ denotes the function from V to $\mathbb{R}$ that takes value 0 for all $w \neq v$ and value 1 for v, which contradicts extremality. Hence the set $\{v \in V : |\boldsymbol{u}(v)| < 1\}$ has at most one element. Considering that $\sum_v \boldsymbol{u}(v) = 0$ gives the result. ■

Appendix II
Proofs of Section IV

A. Convergence of Subgradient Algorithms

Although the result given below is part of the folklore in non-smooth optimization, the convergence proof is often provided with some boundedness assumption of subgradients that are in fact not needed. For the sake of completeness, we provide a self-contained proof.

Assumption 4: Let $\mathbb{H}$ denote a Euclidean space and f a function from $\mathbb{H}$ to $\mathbb{R}$ such that:

1) f is convex continuous with subgradient ∂f.
2) Lower level sets $L_y = \{\boldsymbol{x} \in \mathbb{H} : f(\boldsymbol{x}) \leq \boldsymbol{y}\}$ are bounded.
3) There exists $C > 0$ such that, $\forall \boldsymbol{x} \in \mathbb{H}$, $\forall \boldsymbol{g} \in \partial f(x)$, $\|\boldsymbol{g}\| \leq C(1 + \|\boldsymbol{x}\|)$.

Assumption 5: Let γ_n denote a sequence of positive scalars such that:

1) $\sum_n \gamma_n = +\infty$
2) $\sum_n \gamma_n^2 < +\infty$

Proposition 8: Under Assumptions 4 and 5, any sequence $(\boldsymbol{x}_n)_{n\in\mathbb{N}}$ obeying the subgradient descent scheme:

$$\boldsymbol{x}_{n+1} = \boldsymbol{x}_n - \gamma_n \boldsymbol{g}_n\ ,$$

where $\boldsymbol{g}_n \in \partial f(\boldsymbol{x}_n)$; converges to the set $S = \{\boldsymbol{x} \in \mathbb{H} : f(\boldsymbol{x}) = \inf_{\mathbb{H}} f\}$.

Proof: Since lower level sets are compact (closed by continuity of f and bounded by assumption) there exists a point $\boldsymbol{x}_* \in \mathbb{H}$ such that $f(\boldsymbol{x}_*) = \min_{\boldsymbol{x}\in\mathbb{H}} f(\boldsymbol{x})$. Denote by $\boldsymbol{u}_n = \|\boldsymbol{x}_n - \boldsymbol{x}_*\|^2$. Thus,

$$\boldsymbol{u}_{n+1} = \boldsymbol{u}_n - 2\gamma_n\langle \boldsymbol{x}_n - \boldsymbol{x}_*, \boldsymbol{g}_n\rangle + \gamma_n^2\|\boldsymbol{g}_n\|^2\ .$$

As a consequence of the fundamental property of subgradients, one has $f(\boldsymbol{x}_n) - f(\boldsymbol{x}_*) \leq \langle \boldsymbol{x}_n - x_*, \boldsymbol{g}_n\rangle$. By assumption, $\|\boldsymbol{g}_n\| \leq C(1 + \|\boldsymbol{x}_n\|)$. Hence, there exists some constant $M > 0$ such that:

$$u_{n+1} \leq u_n - 2\gamma_n(f(\boldsymbol{x}_n) - f(\boldsymbol{x}_*)) + \gamma_n^2(M + u_n)$$

Applying Lemma 3 with $\alpha_n = \gamma_n^2$, $v_n = 2\gamma_n(f(\boldsymbol{x}_n) - f(\boldsymbol{x}_*))$ and $\beta_n = \gamma_n^2 M$ yields:

1) u_n converges.
2) $\sum_n v_n < \infty$.

Since $\sum_n \gamma_n = +\infty$ and $f(\boldsymbol{x}_n) - f(\boldsymbol{x}_*) \geq 0$ one necessarily has $\liminf_n f(\boldsymbol{x}_n) - f(\boldsymbol{x}_*) = 0$. Since u_n is bounded, $\boldsymbol{x}_n$ evolves in a compact space and has a convergent subsequence to some point $\tilde{\boldsymbol{x}} \in \mathbb{H}$. By continuity of f, this point $\tilde{\boldsymbol{x}}$ necessarily belongs to the set S. The previous computations are valid with $\boldsymbol{x}_*$ replaced by $\tilde{\boldsymbol{x}}$. Since Lemma 3 ensures that u_n should converge, and it has a subsequence converging to 0, it converges to 0. ■

Lemma 3 (Deterministic Robbins-Siegmund): Assume u_n and v_n are nonnegative scalar sequences, and α_n, β_n are sequences such that: $\sum_n |\alpha_n| < \infty$, $\sum_n |\beta_n| < \infty$ and, for all n,

$$u_{n+1} \leq (1 + \alpha_n)u_n + \beta_n - v_n\ .$$

Then

1) u_n converges to some limit $l \in \mathbb{R}$.
2) $\sum_n v_n < \infty$.

Proof: It is well known that $1 + x \leq \exp x$ for all $x \in \mathbb{R}$. Since u_n and v_n are non-negative, one has: $u_{n+1} \leq \exp(\alpha_n)u_n + \beta_n$. Which iteratively implies: $u_n \leq \exp(\sum_{i=0}^{n-1}\alpha_i)u_0 + \sum_{i=0}^{n-1}\exp(\sum_{j=i+1}^{n-1}\alpha_j)\beta_i$. Since $\sum_n |\alpha_n| < \infty$, for all integers k and l such that $k < l$, $\sum_{n=k}^{l}\alpha_n \leq \sum_{n=0}^{\infty}|\alpha_n|$ and $\exp(\sum_{n=k}^{l}\alpha_n) \leq \exp(\sum_{n=0}^{\infty}|\alpha_n|) < \infty$. Thus, $\sum_n |\beta_n| < \infty$, and it holds that $|u_n| \leq M = \exp(\sum_{n=0}^{\infty}|\alpha_n|)(|u_0| + \sum_{n=0}^{\infty}|\beta_n|) < \infty$. Hence, $|u_{n+1} - u_n| \leq M|\alpha_n| + |\beta_n|$ and $\sum_n |u_{n+1} - u_n| < \infty$. Thence, as stated in the lemma:

1) Sequence $u_n = u_0 + \sum_{m=0}^{n-1}(u_{m+1} - u_m)$ is convergent.
2) $\sum v_n \leq \sum_n |u_{n+1} - u_n| + M\sum_n |\alpha_n| + \sum |\beta_n| < \infty$.

■

B. Derivation of the ADMM

Using the notation introduced in Section IV-B, the update equation of $\boldsymbol{x}_n$ simplifies to:

$$\begin{aligned}\boldsymbol{x}_{n+1}(v) &= \arg\min_x f_v(x) + \sum_{w\sim v} -\boldsymbol{\eta}_n(w,v)x + \frac{\rho}{2}(\boldsymbol{z}_n(w,v) - x)^2 \\ &= \mathrm{prox}_{f_v,\,\rho d(v)}\left(\tilde{\boldsymbol{z}}_n(v) + \frac{\tilde{\boldsymbol{\eta}}_n(v)}{\rho}\right)\end{aligned}$$

where $d(v)$ is the degree of v, $\tilde{\boldsymbol{z}}_n(v) = d(v)^{-1}\sum_{w\sim v}\boldsymbol{z}_n(w,v)$ and $\tilde{\boldsymbol{\eta}}_n(v) = d(v)^{-1}\sum_{w\sim v}\boldsymbol{\eta}_n(w,v)$. Therefore $(\boldsymbol{z}_{n+1}(v,w), \boldsymbol{z}_{n+1}(w,v))^T$ is equal to:

$$\arg\min_{(z_1,z_2)} S_\rho(z_1, z_2; \boldsymbol{\eta}_n(v,w), \boldsymbol{\eta}_n(w,v), \boldsymbol{x}_{n+1}(w), \boldsymbol{x}_{n+1}(v))$$

where $S_\rho(z_1, z_2; \eta_1, \eta_2, x_1, x_2) = \lambda\,|z_1 - z_2| + \eta_1 z_1 + \eta_2 z_2 + \frac{\rho}{2}(z_1 - x_1)^2 + \frac{\rho}{2}(z_2 - x_2)^2$. Minimization of S_ρ w.r.t. (z_1, z_2) yields

$$\frac{z_1 + z_2}{2} = \frac{x_1 + x_2}{2} - \frac{\eta_1 + \eta_2}{2\rho} \tag{18}$$

$$\frac{z_1 - z_2}{2} = \mathrm{soft}_{\lambda/\rho}\left(\frac{x_1 - x_2}{2} + \frac{\eta_2 - \eta_1}{2\rho}\right) \tag{19}$$

where $\mathrm{soft}_\omega(x) = \mathrm{sign}(x)\cdot\max(|x| - \omega, 0)$ is the soft-thresholding function. Using the update equation of $\boldsymbol{\eta}_n$, we obtain

$$\begin{aligned}\boldsymbol{\eta}_{n+1}(v,w) + \boldsymbol{\eta}_{n+1}(w,v) &= \boldsymbol{\eta}_n(v,w) + \boldsymbol{\eta}_n(w,v) \\ &\quad + \rho\,(\boldsymbol{z}_{n+1}(v,w) + \boldsymbol{z}_{n+1}(w,v) - \boldsymbol{x}_{n+1}(v) - \boldsymbol{x}_{n+1}(w))\end{aligned}$$

which by (18) implies that $\boldsymbol{\eta}_{n+1}(v,w) + \boldsymbol{\eta}_{n+1}(w,v) = 0$. Therefore, equation (18) implies that for each n,

$$\boldsymbol{z}_n(v,w) + \boldsymbol{z}_n(w,v) = \boldsymbol{x}_n(v) + \boldsymbol{x}_n(w)$$

We set $\boldsymbol{\delta}_n(v,w) = \frac{1}{2}(\boldsymbol{z}_n(v,w) - \boldsymbol{z}_n(w,v))$. Equation (19) implies that

$$\boldsymbol{\delta}_{n+1}(v,w) = \mathrm{soft}_{\lambda/\rho}\left(\frac{\boldsymbol{x}_{n+1}(w) - \boldsymbol{x}_{n+1}(v)}{2} - \frac{\boldsymbol{\eta}_n(v,w)}{\rho}\right)$$

Using again the update equation of $\boldsymbol{\eta}_n$, we obtain

$$\begin{aligned}\boldsymbol{\eta}_{n+1}(v,w) = \boldsymbol{\eta}_n(v,w) + \rho(&\boldsymbol{\delta}_{n+1}(v,w) \\ &- \frac{1}{2}(\boldsymbol{x}_{n+1}(w) - \boldsymbol{x}_{n+1}(v)))\end{aligned}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2015.2471755, IEEE Transactions on Automatic Control

Set $\boldsymbol{\beta}_n(v,w) = 2\boldsymbol{\eta}_n(v,w)/\rho$. After some algebra,

$$\boldsymbol{\beta}_{n+1}(v,w) = \mathrm{proj}_{[-2\lambda/\rho,2\lambda/\rho]}\left(\boldsymbol{\beta}_n(v,w) + \boldsymbol{x}_{n+1}(v) - \boldsymbol{x}_{n+1}(w)\right) .$$

Finally, we simplify the update equation in $\boldsymbol{x}_n$ as follows. Using $\boldsymbol{z}_n(w,v) = \frac{1}{2}(\boldsymbol{x}_n(w) + \boldsymbol{x}_n(v)) + \boldsymbol{\delta}_n(w,v)$, we obtain $\tilde{\boldsymbol{z}}_n(v) = \frac{1}{2}(\boldsymbol{x}_n(v) + \tilde{\boldsymbol{x}}_n(v)) + \tilde{\boldsymbol{\delta}}_n(v)$ where $\tilde{\boldsymbol{x}}_n(v) = d(v)^{-1}\sum_{w\sim v}\boldsymbol{x}_n(w)$ and $\tilde{\boldsymbol{\delta}}_n(v) = d(v)^{-1}\sum_{w\sim v}\boldsymbol{\delta}_n(w,v)$. By summing equality $2\boldsymbol{\delta}_n(w,v) = \boldsymbol{\beta}_n(w,v) - \boldsymbol{\beta}_{n-1}(w,v) + \boldsymbol{x}_n(v) - \boldsymbol{x}_n(w)$ w.r.t $w \sim v$, one has $2\tilde{\boldsymbol{\delta}}_n(v) = \tilde{\boldsymbol{\beta}}_n(v) - \tilde{\boldsymbol{\beta}}_{n-1}(v) + \boldsymbol{x}_n(v) - \tilde{\boldsymbol{x}}_n(v)$ where $\tilde{\boldsymbol{\beta}}_n(v) = d(v)^{-1}\sum_{w\sim v}\boldsymbol{\beta}_n(w,v)$. Thus, $\tilde{\boldsymbol{z}}_n(v) = \boldsymbol{x}_n(v) + \frac{1}{2}(\tilde{\boldsymbol{\beta}}_n(v) - \tilde{\boldsymbol{\beta}}_{n-1}(v))$. Finally,

$$\boldsymbol{x}_{n+1}(v) = \mathrm{prox}_{f_v,\,\rho d(v)}\left(\boldsymbol{x}_n(v) + \tfrac{3}{2}\tilde{\boldsymbol{\beta}}_n(v) - \tfrac{1}{2}\tilde{\boldsymbol{\beta}}_{n-1}(v)\right) .$$

Setting $\mu_{n+1}(v,w) = \boldsymbol{\beta}_n(v,w)$ yields exactly Algorithm IV-B.

We now assume that the graph G is d-regular and show that the average over the network is preserved. We consider here on the average consensus case. Computing the derivative $\frac{\partial\mathcal{L}}{\partial x}$ which exists since $\mathcal{L}$ is smooth in x; one gets the necessary condition: $\forall v \in V$,

$$\boldsymbol{x}_{n+1}(v) - \boldsymbol{x}_0(v) - \sum_{v\sim w}\boldsymbol{\eta}_n(w,v) + \rho\sum_{v\sim w}\left(\boldsymbol{x}_{n+1}(v) - \boldsymbol{z}_n(w,v)\right) = 0$$

Then, summing up, and using the fact that (i), when an edge (v,w) belongs to $\overleftrightarrow{E}$, the edge (w,v) also belongs to $\overleftrightarrow{E}$, and (ii) $\boldsymbol{\eta}_{n+1}(v,w) = -\boldsymbol{\eta}_{n+1}(w,v)$, along with (iii) $\boldsymbol{z}_n(v,w) + \boldsymbol{z}_n(w,v) = \boldsymbol{x}_n(v) + \boldsymbol{x}_n(w)$:

$$\sum_v (1 + 2\rho d(v))\,\boldsymbol{x}_{n+1}(v) = \sum_v \boldsymbol{x}_0(v) + 2\rho\sum_v d(v)\boldsymbol{x}_n(v)$$

For a d-regular graph, we have, by induction: $\forall n \geq 0$, $\bar{\boldsymbol{x}}_n = \bar{\boldsymbol{x}}_0$.

Appendix III
Proofs of Section V

A. Proof of Proposition 1

Proof: Let $\boldsymbol{u}$ be such that $\|\boldsymbol{u}\|_* \leq 1$ and $\langle \boldsymbol{u}, \boldsymbol{x}\rangle = \|\boldsymbol{x}\|_{\mathrm{TV}}$. Then, $\langle \boldsymbol{u}, \boldsymbol{y} - \boldsymbol{x}\rangle + \|\boldsymbol{x}\|_{\mathrm{TV}} = \langle \boldsymbol{u}, \boldsymbol{y}\rangle \leq \|\boldsymbol{u}\|_*\|\boldsymbol{y}\|_{\mathrm{TV}} = \|\boldsymbol{y}\|_{\mathrm{TV}}$ which means that $\boldsymbol{u} \in \partial\|\boldsymbol{x}\|_{\mathrm{TV}}$. Conversely, assume $\boldsymbol{u} \in \partial\|\boldsymbol{x}\|_{\mathrm{TV}}$ and $\boldsymbol{x}_u$ is s.t. $\|\boldsymbol{x}_u\|_{\mathrm{TV}} = 1$ and $\langle \boldsymbol{u}, \boldsymbol{x}_u\rangle = \|\boldsymbol{u}\|_*$. Define $\boldsymbol{y}_u = \|\boldsymbol{x}\|_{\mathrm{TV}}\boldsymbol{x}_u$; one has: $\|\boldsymbol{y}_u\|_{\mathrm{TV}} - \|\boldsymbol{x}\|_{\mathrm{TV}} \geq \langle \boldsymbol{u}, \boldsymbol{y}_u - \boldsymbol{x}\rangle$, which gives $0 \geq \|\boldsymbol{u}\|_*\|\boldsymbol{x}\|_{\mathrm{TV}} - \langle \boldsymbol{u}, \boldsymbol{x}\rangle$. By inequality $\|\boldsymbol{u}\|_*\|\boldsymbol{x}\|_{\mathrm{TV}} - \langle \boldsymbol{u}, \boldsymbol{x}\rangle \geq 0$, one has $\langle \boldsymbol{u}, \boldsymbol{x}\rangle = \|\boldsymbol{u}\|_*\|\boldsymbol{x}\|_{\mathrm{TV}}$. Moreover, as $\boldsymbol{u} \in \partial\|\boldsymbol{x}\|_{\mathrm{TV}}$ $\|2\boldsymbol{x}\|_{\mathrm{TV}} - \|\boldsymbol{x}\|_{\mathrm{TV}} \geq \langle \boldsymbol{u}, \boldsymbol{x}\rangle = \|\boldsymbol{u}\|_*\|\boldsymbol{x}\|_{\mathrm{TV}}$. Consequently, f $\boldsymbol{x} \neq 0$, then $\|\boldsymbol{u}\|_* = 1$.

If $\boldsymbol{x} = 0$, then writing $\langle \boldsymbol{u}, \boldsymbol{x}_u\rangle \leq \|\boldsymbol{x}_u\|_{\mathrm{TV}}$ directly leads to $\|\boldsymbol{u}\|_* \leq 1$. ■

B. Proof of Lemma 1

Proof: First notice that the integral is well defined since $\lambda \in \mathbb{R} \mapsto \mathrm{Per}(\{\boldsymbol{x} \geq \lambda\})$ has its support included in $[\min \boldsymbol{x}, \max \boldsymbol{x}]$ (recall that $\boldsymbol{x}$ is seen as a function) and is piecewise constant with finite values. For each edge $\{v,w\} \in E$, denote by $I_e = [\boldsymbol{x}(v) \wedge \boldsymbol{x}(w), \boldsymbol{x}(v) \vee \boldsymbol{x}(w)] \subset \mathbb{R}$. Now, it is easy to check that $\mathrm{Per}(\{\boldsymbol{x} \geq \lambda\}) = \sum_{e\in E}\mathbf{1}_{I_e}(\lambda)$. Hence,

$$\int_{-\infty}^{+\infty}\mathrm{Per}(\{\boldsymbol{x} \geq \lambda\})\mathrm{d}\lambda = \sum_{e\in E}\int_{-\infty}^{+\infty}\mathbf{1}_{I_e}(\lambda)\mathrm{d}\lambda = \sum_{e\in E}|I_e|$$

where $|I|$ denotes the length $b-a$ of interval $I = [a,b]$. The rightmost term is equal to $\|\boldsymbol{x}\|_{\mathrm{TV}}$, which completes the proof. ■

C. Proof of Proposition 4

Proof: Define $\mathrm{grad}\,\boldsymbol{x} : (v,w) \in \vec{E} \mapsto \boldsymbol{x}(w) - \boldsymbol{x}(v) \in \mathbb{R}$. First the set $\Xi_u = \{\boldsymbol{\xi} \in \mathbb{R}^{\vec{E}} : \boldsymbol{u} = \mathrm{div}\,\boldsymbol{\xi}\}$ is not empty. Indeed, grad : $\mathbb{R}_0^V \to \mathbb{R}^{\vec{E}}$ is into since G is connected. Hence div : $\mathbb{R}^{\vec{E}} \to \mathbb{R}_0^V$ is onto.

Assume $\boldsymbol{u} = \mathrm{div}\,\boldsymbol{\xi}$, then $\langle \boldsymbol{u}, \boldsymbol{x}\rangle = -\langle \boldsymbol{\xi}, \mathrm{grad}\,\boldsymbol{x}\rangle \leq \|\boldsymbol{\xi}\|_\infty\|\boldsymbol{x}\|_{\mathrm{TV}}$. Hence $\|\boldsymbol{u}\|_* \leq \|\boldsymbol{\xi}\|_\infty$, and $\|\boldsymbol{u}\|_* \leq \inf\{\|\boldsymbol{\xi}\|_\infty : \boldsymbol{u} = \mathrm{div}\,\boldsymbol{\xi}\}$. Reciprocally, since G is assumed to be connected, grad : $\mathbb{R}_0^V \to \mathbb{R}^{\vec{E}}$ is into; let us denote by R its range. Then grad : $(\mathbb{R}_0^V, \|\cdot\|_{\mathrm{TV}}) \to (R, \|\cdot\|_1)$ is an isometry. Moreover, using Hahn-Banach theorem (finite dimensional case), one can embed isometrically $(R, \|\cdot\|_1)^*$ into $(\mathbb{R}^{\vec{E}}, \|\cdot\|_1)^* \simeq (\mathbb{R}^{\vec{E}}, \|\cdot\|_\infty)$. Hence to $\boldsymbol{u} \in (\mathbb{R}_0^V, \|\cdot\|_*)$ one can associate $\boldsymbol{\varphi} \in \mathbb{R}^{\vec{E}}$ such that $\langle \boldsymbol{u}, \boldsymbol{x}\rangle = \langle \boldsymbol{\varphi}, \mathrm{grad}\,\boldsymbol{x}\rangle$ with $\|\boldsymbol{\varphi}\|_\infty = \|\boldsymbol{u}\|_*$. It suffices to take $\boldsymbol{\xi} = -\boldsymbol{\varphi}$ to have $\boldsymbol{u} = \mathrm{div}\,\boldsymbol{\xi}$ with $\|\boldsymbol{u}\|_* = \|\boldsymbol{\xi}\|_\infty$ to prove the reverse inequality. ■

D. Proof of Proposition 5

Proof:

1) Proof of equality $\|\boldsymbol{u}\|_* = \max_{\emptyset\subsetneq S\subsetneq V}\frac{\langle \boldsymbol{u}, \mathbf{1}_S\rangle}{\|\mathbf{1}_S\|_{\mathrm{TV}}}$:

Since $\mathbb{R}_0^V$ is finite dimensional, there exists $\boldsymbol{x}_u \in \mathbb{R}_0^V$ with $\|\boldsymbol{x}_u\|_{\mathrm{TV}} = 1$ such that $\|\boldsymbol{u}\|_* = \langle \boldsymbol{u}, \boldsymbol{x}_u\rangle$. Since $\langle \boldsymbol{u}, \mathbf{1}_V\rangle = 0$ one has $\langle \boldsymbol{u}, \tilde{\boldsymbol{x}}_u\rangle = \langle \boldsymbol{u}, \boldsymbol{x}_u\rangle$ with $\tilde{\boldsymbol{x}}_u = \boldsymbol{x}_u - (\min_v \boldsymbol{x}_u(v))\mathbf{1}_V$. Now, let us consider subsets of V having the form $S_\mu = \{\tilde{\boldsymbol{x}}_u \geq \mu\}$ for $\mu \in \mathbb{R}$. Notice that $S_\mu = V$ for $\mu \leq 0$ and $S_\mu = \emptyset$ for $\mu > M$ with $M > 0$ large enough. Hence, the following integral is well defined:

$$\int_{-\infty}^{+\infty}\langle \boldsymbol{u}, \mathbf{1}_{S_\mu}\rangle\mathrm{d}\mu$$

And,

$$\int_{-\infty}^{+\infty}\langle \boldsymbol{u}, \mathbf{1}_{S_\mu}\rangle\mathrm{d}\mu = \int_0^M\langle \boldsymbol{u}, \mathbf{1}_{S_\mu}\rangle\mathrm{d}\mu = \langle \boldsymbol{u}, \int_0^M \mathbf{1}_{S_\mu}\mathrm{d}\mu\rangle .$$

where $\int_0^M \mathbf{1}_{S_\mu}\mathrm{d}\mu$ denotes function $v \mapsto \int_0^M \mathbf{1}_{S_\mu}(v)\mathrm{d}\mu$. Moreover $\forall v \in V$,

$$\int_0^M \mathbf{1}_{S_\mu}(v)\mathrm{d}\mu = \int_0^{+\infty}\mathbf{1}_{\{\mu\leq\tilde{\boldsymbol{x}}_u(v)\}}\mathrm{d}\mu = \tilde{\boldsymbol{x}}_u(v)$$

Hence

$$\int_{-\infty}^{+\infty}\langle \boldsymbol{u}, \mathbf{1}_{S_\mu}\rangle\mathrm{d}\mu = \langle \boldsymbol{u}, \tilde{\boldsymbol{x}}_u\rangle .$$

By definition of $\|\boldsymbol{u}\|_*$ and the fact that (i) $\langle \boldsymbol{u}, \mathbf{1}_V\rangle = 0$, (ii) $\|\boldsymbol{x}\|_{\mathrm{TV}} = \|\boldsymbol{x} + c\mathbf{1}_V\|_{\mathrm{TV}}$, one has $\langle \boldsymbol{u}, \mathbf{1}_{S_\mu}\rangle \leq \|\boldsymbol{u}\|_*\|\mathbf{1}_{S_\mu}\|_{\mathrm{TV}}$. Hence, function $\mu \mapsto \langle \boldsymbol{u}, \mathbf{1}_{S_\mu}\rangle - \|\boldsymbol{u}\|_*\|\mathbf{1}_{S_\mu}\|_{\mathrm{TV}}$ is nonpositive. Integrating and using the co-area formula, one gets, for almost every $\mu \in \mathbb{R}$:

$$\langle \boldsymbol{u}, \mathbf{1}_{S_\mu}\rangle = \|\boldsymbol{u}\|_*\|\mathbf{1}_{S_\mu}\|_{\mathrm{TV}}$$

A fortiori, the set of such μ is not empty, which concludes the proof of the first equality.

2) Proof of equality

$$\max_{\emptyset\subsetneq S\subsetneq V}\frac{\langle \boldsymbol{u}, \mathbf{1}_S\rangle}{\|\mathbf{1}_S\|_{\mathrm{TV}}} = \max_{\substack{\emptyset\subsetneq S\subset V, |S|\leq|V|/2 \\ G(S)\ \textit{is connected}}}\frac{|\langle \boldsymbol{u}, \mathbf{1}_S\rangle|}{\|\mathbf{1}_S\|_{\mathrm{TV}}}$$

:

Using the fact that $\boldsymbol{u} \in \mathbb{R}_0^V$, we get that $|\langle \boldsymbol{u}, \mathbf{1}_S\rangle| = |\langle \boldsymbol{u}, \mathbf{1}_{V\setminus S}\rangle|$. Moreover, using the first equality of Proposition 5 and equality $\mathrm{Per}(S) = \mathrm{Per}(V \setminus S)$, we get that $\|\boldsymbol{u}\|_* = \max_{\emptyset\subsetneq S\subset V, |S|\leq|V|/2}\frac{|\langle \boldsymbol{u}, \mathbf{1}_S\rangle|}{\|\mathbf{1}_S\|_{\mathrm{TV}}}$. Let us consider a subset S such that $\|\boldsymbol{u}\|_* = \frac{|\langle \boldsymbol{u}, \mathbf{1}_S\rangle|}{\|\mathbf{1}_S\|_{\mathrm{TV}}}$ and assume that $S = S_1 \cup S_2$ with $E(S_1, S_2) = \emptyset$ where $E(S_1, S_2)$ is the set of edges having one extremity in S_1 and one extremity in S_2. Then the ratio $\frac{|\langle \boldsymbol{u}, \mathbf{1}_S\rangle|}{\|\mathbf{1}_S\|_{\mathrm{TV}}}$ can be written as $\frac{|\langle \boldsymbol{u}, \mathbf{1}_{S_1}\rangle + \langle \boldsymbol{u}, \mathbf{1}_{S_2}\rangle|}{\|\mathbf{1}_{S_1}\|_{\mathrm{TV}} + \|\mathbf{1}_{S_2}\|_{\mathrm{TV}}}$ which is less than or equal to $\frac{|\langle \boldsymbol{u}, \mathbf{1}_{S_1}\rangle| + |\langle \boldsymbol{u}, \mathbf{1}_{S_2}\rangle|}{\|\mathbf{1}_{S_1}\|_{\mathrm{TV}} + \|\mathbf{1}_{S_2}\|_{\mathrm{TV}}}$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TAC.2015.2471755, IEEE Transactions on Automatic Control

The last ratio is clearly bounded by $\max\left(\frac{|\langle \boldsymbol{u}, \mathbf{1}_{S_1}\rangle|}{\|\mathbf{1}_{S_1}\|_{\mathrm{TV}}}, \frac{|\langle \boldsymbol{u}, \mathbf{1}_{S_2}\rangle|}{\|\mathbf{1}_{S_2}\|_{\mathrm{TV}}}\right)$. Since S is maximizing the ratio $\frac{|\langle \boldsymbol{u}, \mathbf{1}_{S}\rangle|}{\|\mathbf{1}_{S}\|_{\mathrm{TV}}}$, we should have $\frac{|\langle \boldsymbol{u}, \mathbf{1}_{S}\rangle|}{\|\mathbf{1}_{S}\|_{\mathrm{TV}}} = \max\left(\frac{|\langle \boldsymbol{u}, \mathbf{1}_{S_1}\rangle|}{\|\mathbf{1}_{S_1}\|_{\mathrm{TV}}}, \frac{|\langle \boldsymbol{u}, \mathbf{1}_{S_2}\rangle|}{\|\mathbf{1}_{S_2}\|_{\mathrm{TV}}}\right)$. Consequently, to compute $\|\boldsymbol{u}\|_*$, we can focus on subsets S inducing a connected graph $G(S)$. ∎

E. Proof of Proposition 6

Proof: First, observe that since $\langle \boldsymbol{u}, \mathbf{1}_{S_{i-1}}\rangle - \lambda_i \mathrm{Per}(S_{i-1}) = 0$, and $S_i = \arg\max_{S\subset V} \langle \boldsymbol{u}, \mathbf{1}_S\rangle - \lambda_i \mathrm{Per}(S)$, we necessarily have $\langle \boldsymbol{u}, \mathbf{1}_{S_i}\rangle - \lambda_i \mathrm{Per}(S_i) \geq 0$. In other words, we have $\frac{\langle \boldsymbol{u}, \mathbf{1}_{S_i}\rangle}{\mathrm{Per}(S_i)} \equiv \lambda_{i+1} \geq \lambda_i$.

Let us now assume that we obtained $\lambda_{i+1} = \lambda_i$ for some i. This is equivalent to say that $\langle \boldsymbol{u}, \mathbf{1}_{S_i}\rangle - \lambda_i \mathrm{Per}(S_i) = 0$. By definition of S_i, we should have $\langle \boldsymbol{u}, \mathbf{1}_S\rangle - \lambda_i \mathrm{Per}(S) \leq \langle \boldsymbol{u}, \mathbf{1}_{S_i}\rangle - \lambda_i \mathrm{Per}(S_i)$ for each subset. Consequently, $\frac{\langle \boldsymbol{u}, \mathbf{1}_S\rangle}{\mathrm{Per}(S)} \leq \lambda_i$ for each S with equality for $S = S_{i-1}$. The convergence of the algorithm is then achieved.

To finish the proof, let us try to bound the number of iterations of the algorithm. Assume that $\lambda_{i+1} > \lambda_i$. This implies that $\langle \boldsymbol{u}, \mathbf{1}_{S_i}\rangle - \lambda_i \mathrm{Per}(S_i) > 0 = \langle \boldsymbol{u}, \mathbf{1}_{S_{i-1}}\rangle - \lambda_i \mathrm{Per}(S_{i-1})$. Moreover, by definition of S_{i-1}, we should have $\langle \boldsymbol{u}, \mathbf{1}_{S_i}\rangle - \lambda_{i-1} \mathrm{Per}(S_i) \leq \langle \boldsymbol{u}, \mathbf{1}_{S_{i-1}}\rangle - \lambda_{i-1} \mathrm{Per}(S_{i-1})$. These two inequalities can be written as follows: $\lambda_i(\mathrm{Per}(S_i) - \mathrm{Per}(S_{i-1})) < \langle \boldsymbol{u}, \mathbf{1}_{S_i}\rangle - \langle \boldsymbol{u}, \mathbf{1}_{S_{i-1}}\rangle \leq \lambda_{i-1}(\mathrm{Per}(S_i) - \mathrm{Per}(S_{i-1}))$. This holds only if $\mathrm{Per}(S_i) - \mathrm{Per}(S_{i-1}) < 0$. Said another way, if $\lambda_{i+1} > \lambda_i$ then $\mathrm{Per}(S_i)$ decreases at least by one unit. Since we know that $\mathrm{Per}(S_i)$ is between 1 and $|E| - 1$, the number of iterations is bounded by $|E|$. ∎

Appendix IV
Proofs of Section VI

A. Proof of Theorem 4

We are going to use Theorem 1 where $f_v(x)$ is given by $\frac{1}{2}(x - \boldsymbol{x}_0(v))^2 + \lambda \sum_{w\in S} |x - \boldsymbol{x}_0(w)|$.

Let us first consider the case $\overline{x}_0^R + \lambda|S| < a$. The subdifferential $\partial F(x^\star \mathbf{1}_V)$ for $x^* = \overline{x}_0^R + \lambda|S|$ is clearly given by the singleton $\{x^* \mathbf{1}_R - \boldsymbol{x}_0^R - \lambda|S|\mathbf{1}_R\} = \{\overline{x}_0^R \mathbf{1}_R - \boldsymbol{x}_0^R\}$. Since $\lambda \geq ||\boldsymbol{x}_0^R - \overline{x}_0^R \mathbf{1}_R||_*$, the second condition of Theorem 1 is satisfied implying that $x^* \mathbf{1}_V$ is a minimizer of (17). The strict convexity of (17) implies the uniqueness of the minimizer.

The case $\overline{x}_0^R - \lambda|S| > a$ is exactly handled in the same way.

Let us now assume that $|\overline{x}_0^R - a| \leq \lambda|S|$ and take $x^* = a$. Using the fact that the subdifferential of $|x|$ for $x = 0$ is given by the interval $[-1, 1]$, we deduce that

$$\partial f_v(x^\star = a) = \lambda \times \{ \sum_{\substack{w\in S \\ |\zeta_{v,w}|\leq 1}} \zeta_{v,w} \} + a - \boldsymbol{x}_0(v).$$

Let $\zeta = \frac{\overline{x}_0^R - a}{\lambda |S|}$ and let us take $\zeta_{v,w} = \zeta$ for $v \in R$ and $w \in S$. $\zeta_{v,w}$ clearly belongs to the interval $[-1, 1]$. Moreover, using the fact that $\partial F(x^\star \mathbf{1}_R) = \left\{u \in \mathbb{R}^R : \forall v \in R,\, u(v) \in \partial f_v(x^\star)\right\}$, we deduce that $\partial F(x^\star \mathbf{1}_V)$ contains the vector $x^* \mathbf{1}_R - \boldsymbol{x}_0^R + \lambda\zeta|S|\mathbf{1}_R$. Using the definition of ζ, we get that $\overline{x}_0^R \mathbf{1}_R - \boldsymbol{x}_0^R$ belongs to $\partial F(x^\star \mathbf{1}_V)$. Using again the assumption $\lambda \geq ||\boldsymbol{x}_0^R - \overline{x}_0^R \mathbf{1}_R||_*$ and Theorem 1 we get that $x^* \mathbf{1}_V$ is a minimizer of (17). Uniqueness of the minimizer is still implied by the strict convexity of (17).

Walid Ben-Ameur has obtained an engineering degree, a master degree and a Ph.D. from Télécom Paristech, France. He was with Orange Labs from 1999 to 2001. He is currently a full professor in Télécom SudParis and head of the Laboratory Samovar, CNRS UMR 5157. His main interests are combinatorial optimization, mathematical programming and network Optimization. He was a winner in 2006 of the Robert Faure prize and a co-winner of the Glover-Klingman prize in 2007.

Pascal Bianchi was born in 1977 in Nancy, France. He received the M.Sc. degree of the University of Paris XI and Supélec in 2000 and the Ph.D. degree of the University of Marne-la-Vallée in 2003. From 2003 to 2009, he was an Associate Professor at the Telecommunication Department of Supélec. In 2009, he joined the Statistics and Applications group at LTCI-Telecom ParisTech. His current research interests are in the area of numerical optimization, signal processing and distributed systems.

Jérémie Jakubowicz received the MS degree (2004) and the PhD degree (2007) in applied mathematics from the Ecole Normale Supérieure de Cachan. From 2008 to 2011, he was an Assistant Professor at Télécom ParisTech. Since 2011 he has been an Associate Professor at Télécom SudParis and an Associate Researcher with the CNRS. His current research interests include distributed statistical signal processing, image processing and data mining.