Correlation among smart grid actors

Nicolas Gensollen, Monique Becker, Vincent Gauthier and Michel Marot
CNRS UMR 5157 SAMOVAR,
Telecom SudParis/Institut Mines Telecom
Email: {nicolas.gensollen, vincent.gauthier, michel.marot, monique.becker}@telecom-sudparis.eu

***Abstract*—**

I. Introduction

Designing stable power systems is a classical engineering challenge since blackout events can have catastrophic consequences. Nevertheless, stability is a general concept and can be studied from several points of view. A recent approach introduced by complex systems theorists consists in abstracting the power grid as a graph where nodes represent loads, generators, or transformers, while edges stand for electrical lines. As graphs, power grids display some characteristics like clustering coefficients, degree distributions, or density for instance. These metrics enable scientists to get a better understanding of the structure of such complex systems. It appears that power grids are particular graphs compared to well-known large networks like the internet, the world wide web, or social networks. That is, they exhibit different characteritics, suggesting that they tend to behave differently. A common hypothesis for such a difference is that power grids are completely designed by a very small number of entities (electricity operators mainly), whereas other pre-cited graphs results from the interactions of billions of independant entities. Therefore, meaningfull properties like power law degree distributions, prerential attachment, or hub oriented attacks do not seem to apply with power grids.

Actually, drawing conclusions based only on the study of the underlying graph structure may lead to unrealistic results. Power grids are ideed dynamical systems governed by physical laws. Concepts such as shortest path or centrality metrics have thus to be extended to take into account the particular nature of power grids. Besides, stability and resilience of such systems are much more complex than the simple connectivity of the underlying graph. It is well-known for instance that frequency synchronization in the grid is a necessary condition for stability. A common solution consists in modeling the grid as a network of coupled oscillators governed by a swing equation dynamic. It can then be shown that the topology of the network impacts directly the capacity of the oscillators to reach the synchronized state.

Understanding power grids seems like a reasonable strarting point for studying so-called *smart grids* that gained a lot of attention recently. As there exists many good articles presenting smart grid ideas, concepts, or architectures, we will not dwell too much on this point. In this paper, we rather concentrate on a different approach to the system stability. We consider a set of agents that are equiped with distributed energy ressources (DER) as well as electrical loads, such that an agent can produce and consume energy at any time. His production can be, of course, used to meet his own demand, but in cases where he is over-producing, we consider that he has the possibility to sell his extra-production to the grid. Such an agent model is known as a "*prosumer*" (and will be called accordingly in this paper).

The constraints necessary for the electrical grid to remain in a stable state obiously forbid any anarchic system where prosumers inject power whenever and however they want. A solution for organizing such a system consists in a market environement, where participating entities announce a production capacity for an upcoming period of time. These entities thus commit to injecting exactly at any time the contracted amount of power under financial penalities if they fail. Such a market system necessitates obviously some communication architecture to be viable. Since the number of prosumers in a system may be large, it seems unrealistic to have a flat centralized topology where a single agent coordinates all the others individually. Moreover, taking into account that the production component of these prosumers is supposed to come mainly from renewables, the over-producing state of the prosumers, and therefore the power injected in the grid, might be rather unstable. Imagine indeed how strong the impact of a simple cloud on a single agent production will be.

It has been suggested that energy diversification, geographical expansion, as well as using storage devices are envisageable techniques for stabilizing the production. Considering virtual coalitions of prosumers as potential providers, and allowing these aggregations to enter the market, could potentially solve these problems. Indeed, the system operator would only have to communicate with a relatively small number of aggregators, that in turn will use some internal communication protocol to manage the involved agents. Moreover, if the aggregation has been done properly, one expect a more stable and predictable energy production for the coalition than for a single agent. This diversification idea is ideed a central topic of the present paper : given N prosumers, what coalitions should be formed so that the compromise between expected production and variability is optimized ?

We will see that variability in the coalitions productions can be quantified to a certain extend by the correlation among the agents forming the coalitions. Understanding the correlation reliationships among the agents can thus be illuminating in the sense that it will help us decide both what coalitions to form and how much they should sell. More precisely, we will build a framework in which the system operator has the possibility of

fixing some entrance conditions on the market both in terms of stability and sufficient production. Based on predictions, coalitions can decide wether or not to enter and what power quantity they are willing to provide. Of course, coalitions failing at fulfilling their obligations during the contracted period of time (because of bad forecasts, unexpected rare events, or lying) will be exposed to financial penalities.

The paper is organized as follows, section 2 will give a brief overview of the related litterature, section 3 will clarify how we generated realistic prosumer production traces based on weather data. In section 4, we will define most of the notations and explain why correlation between prosumer is a quantity of interest for our objective. Based on the conclusions of section 4, section 5 shows how we approached the problem in a complex system fashion. Finally, section 6 will provide some results both on performance of the method and resilience of the coalitions formed.

II. Related Work

The intermittent nature of renewables such as wind or solar power introduces new challenges in control design. Besides, on the contrary to fossil plants whose production can be scheduled in advance to meet the expected consumption, renewables by definition only produce when the natural ressource is present. Unfortunately, these moments do not necessarily coincide with the consumption peak hours. A first approach to remedy this situation by acting on the consumption component is the use of dynamic ectricity prices sent to the end users. Together with demand side management techniques implemented on the smart meters, they consitute a first tool for shaping the consumption profiles. In periods where the production is expected to be greater than the consumption, prices are scheduled low in order to encourage end users to delay their elastic loads to these periods. On the contrary, when the production is expected to be less than the consumption, high prices are scheduled to deter any delayable loads over these periods.

Another popular approach on the production side, is to combined renewable generators with storage devices, such that these are charged when there is a surplus of production, and discharged when the consumption exceeds the production. Although simple, this idea causes numerous challenges. Depending on the size of the system considered, centralized or decentralized control is desirable. In[], the authors introduce a distributed energy management system with a high penetration of renewables such that power is scheduled in a distributed fashion. In [] the optimal storage capacity problem is adressed. There is indeed an interesting tradeoff between the costs of the equipments and the expected availability of power. The authors develop a framework that enables them to exhibit a Pareto front of efficient solutions.

Although there are great progresses in stabilizing the production by demand side managment or storage architectures, the need for good prediction techniques is more important than ever. Understanding the statistical properties of wind or solar irradiance are already well-established areas of research. There exists indeed a very interesting litterature on techniques aiming at predicting wind or solar irradiance profiles for future periods with error margins. In this paper, we adopt a slightly different approach : we suppose that the agents have historical records of their production and consumption, and we seek aggregations of agents with high expected production and low variability (wich is sometimes called risk in the present paper).

Optimization of expected returns to risk is a traditional goal in finance. It is indeed well-known, that the more risk one is willing to take, the higher his potential gains. On the contrary, when investing exclusively on low rik assets, one should expect relatively small gains. This tradeoff is formalized in the Markowitz' portfolio theory. More precisely, given a set of assets for which we have some historic data of returns, the objective is to find a linear combination of these assets (the so-called portfolio) which maximizes the expected value while minimizing the variance of the portfolio's return. Markowitz's answer is a set of efficient portolios that all optimize in some sense this tradeoff. If one is able to put a number on his risk acceptance or on the target expected return, the corresponding efficient portfolio is a priori the best option.

This problem exhibits similarities with our objective of forming stable coalitions since our expected return can be formulated as the expected production and the risk as the variance of the production. However, one of the assumptions in the portfolio theory is that returns are normaly distributed random variables (or at least that the joint distribution is elliptic). This is a strong assumption which reveals unrealistic in most cases for our purposes when looking at real data.

Nevertheless, one of the key point in the Markowitz theory is to consider explicitely the correlation structure between the assets since these correlation relationships impact directly (as we will see in section X) the variances of the portfolios. Since the work of Mantanega, an interresting approach consists in computing a distance metric based on the correlation coefficients in order to organize the series in a correlation graph where the weight of an edge between two timeseries (nodes) is the metric value for these series.

Because the metric can be computed for all pairs, these graphs are complete and of little use as is. Historically, the approach used by Mantanega was to compute a minimum spanning tree over the correlation graph as to extract a correlation structure of the form of a hierarchical clustering. Later on, it was pointed out that, by definition, a spanning tree could not capture the underlying clustering structure hidden in the correlation graph. Another filtering approach by mean of a threshold ϵ on the edge weights solves this problem and allows one to exhibit clusters of correlated series. In the following we will refer to these filtered graphs as ϵ-graph.

An efficient portfolio of assets and a clustering of assets based on the correlations are obviously two very distinct things. And none of them is exactly what we seek. We will see in section X, how we combined both theories in order to form sufficiently producing coalitions with an acceptable risk level.

III. Generating realistic prosumer patterns

An essential component of the smart grid is the smart meter which makes the interface between the end user and the rest of the system. Smart meters coupled with sensors measure quantities of interest like instantaneous consumption, receive informations from the grid (electricity price for instance), and take actions accordingly (demand side management program). Smart meters are currently and gradually deployed, and will probably provide interesting datasets to work on. Unfortunately, at the time this paper was written, production and consumption data for prosumers over a large region were not yet available to our knowledge. Some interesting experiments are notwithstanding being conducted and data are progressively made public (see for instance the ISSDA experiment in Ireland).

We denote by $P_i(t)$ the instantaneous extra-production of agent i at time t. That is, $P_i(t) = P_i^P(t) - P_i^D(t)$, where $P_i^P(t)$ represents the total production of agent i at time t and $P_i^D(t)$ its consumption at time t. In other words, $P_i(t)$ represents the instantaneous surplus of power that agent i is willing to sell at time t. As explained above, since large datasets containing this quantity over time are not yet available, we simulated these traces by considering separately P_i^P and P_i^D.

For a prosumer i, it is possible to write both quantities as a sum over the distributed energy ressources (DER_i) and loads ($load_i$) of i :

$$P_i^P(t) = \sum_{k \in DER_i} P_k(t) \tag{1}$$

$$P_i^D(t) = \sum_{k \in load_i} P_k(t) \tag{2}$$

For simplicity, in this paper we only consider windturbines (WT) and photovoltaic panels (PV) as possible DER for the agents ($DER_i = WT_i \cup PV_i$):

$$P_i^D(t) = \sum_{k \in WT_i} P_k(t) + \sum_{k \in PV_i} P_k(t) \tag{3}$$

We denote by $\nu_i(t)$ and $\xi_i(t)$ the wind speed (in $m.s^{-1}$) and the solar irradiance (in $W.m^{-2}$) at the location of agent i and at time t, so that :

$$P_i^P(t) = \sum_{k \in WT_i} \mathcal{F}_{WT}(\nu_i(t)) + \sum_{k \in PV_i} \mathcal{F}_{PV}(\xi_i(t)) \tag{4}$$

Where $\mathcal{F}_{WT}$ (resp. $\mathcal{F}_{PV}$) is the power curve for the windturbines (resp. photovoltaic panels). We made here the implicit assumption that all windturbines (resp. photovoltaic panels) have the same power curve. The model can be easily extended to muplitiple power curves accounting for different types of generators. More details about power curves and their approximations can be found in [].

Weather quantities like wind speed or solar irradiance appear thus as alternative data for generating the P_i series. Fortunately, these kind of data are easier to find, and since the developement of small personal weather stations, their geographical granularity keeps increasing. A quantity like wind speed depends indeed both on time and location. In the following, time is discretized into slots and space into zones. A zone is simply a portion of the considered space for which we have weather data. Therefore, if prosumers i and j are positioned on the same zone, they are exposed to the same weather. Adding some noise can easily be done though not considered in this paper. The process for generating the P_i series is pictured in the first block of the process diagram (see figure 1).

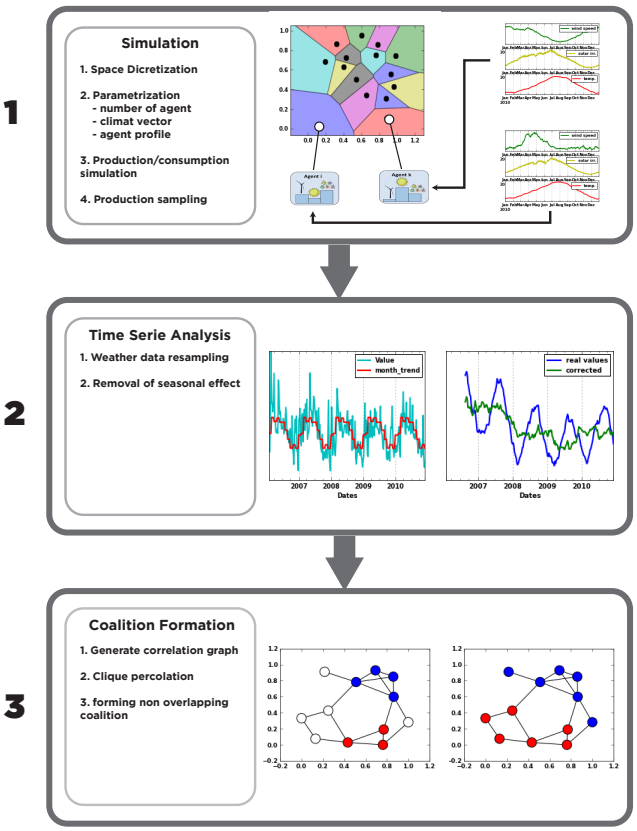

Figure 1: Process diagram

Note that a prosumer i is defined by his zone Z_i as well as the sets DER_i and $load_i$. That is, a prosumer can be configured to represent anything from a single windturbine for instance ($DER_i = \{WT_0\}$ and $load_i = \emptyset$) to a pure load ($DER_i = \emptyset$ and $load_i = \{L_0\}$) throug more complex combinations. In practice, we use random configurations for the agents.

In the rest of the paper, we use french weather data (see www.infoclimat.fr) starting in january 2006 and ending in december 2012, with a sampling frequency of three hours, and generate N timeseries of extra-production over this date range.

IV. Notations

As explained in section III, we have a set $\mathcal{A} = \{a_1, a_2, ..., a_N\}$ of N prosumers configured randomly, and for each agent, we simulated his extra-production $P_i(t),\ \forall i \in \mathcal{A}$ from 2006 to 2012. Based on these historical values, our objective is now to form groups of prosumers (the so-called coalitions) so that the global power production resulting from the superposition of individual's extra-productions be both sufficiently high and predictable. Let $P_S(t) = \sum_{i \in S} P_i(t)$ be the extra-production of coalition S at time t.

Suppose now that coalition S has to suggest a production value P_S^{CRCT} to enter the market. This means that, during the time S is on the market, it will have to inject in the grid exactly P_S^{CRCT} at any time t and will be rewarded proportionally to this amount, with penalities if it deviates. Obviously, the actual extra-production will not be constant at this value and will oscillate due to intermittencies in the production and consumption. If S always produces more than P_S^{CRCT}, it will never have to pay penalities, but it is losing some gains since it could have annonced a higher contract value. If the production oscillates around P_S^{CRCT}, by using batteries or demand side management techniques (see section II), S could be able to maintain its production to the contract value at any time. Nevertheless, if the oscillations are too important compared to the available storage capacity, S will probably break the contract and pay penalities. We can see that there is a return over risk tradeoff here, meaning that coalitions should find the right balance between announcing too low and losing some potential gains, and claiming too high and paying penalities.

Let us illustrate the rest of the notations and concepts with a simple example. We consider only two agents i and j such that the distribution of their extra-production can be approximated by normal distributions : $P_i \sim \mathcal{N}(\mu_i, \sigma_i)$ and $P_j \sim \mathcal{N}(\mu_j, \sigma_j)$. This is only for explanation purposes as it is of course rather unrealistic in real situations where the distributions are skewed. Using simple statistics, we can write the distribution of the coalition $\{i, j\}$ as $P_{\{i,j\}} \sim \mathcal{N}(\mu_{ij}, \sigma_{ij})$, where :

$$\begin{cases} \mu_{ij} = \mu_i + \mu_j \\ \sigma_{ij} = \sqrt{\sigma_i^2 + \sigma_j^2 + \rho_{ij}\sigma_i\sigma_j} \end{cases} \tag{5}$$

ρ_{ij} being the Pearson's correlation coefficient between P_i and P_j. If the coalition $\{i, j\}$ proposes a contract value P_S^{CRCT}, all instants where $\{i, j\}$ will produce less than P_S^{CRCT} is critical. Indeed, in this situation, $\{i, j\}$ will either have to discharge batteries to keep up with its contract, or pay penalities to the grid. The probability that $\{i, j\}$ is underproducing compared to the contract : $Pr[P_{i,j} \leq P^{CRCT}]$ is thus an important indicator of the coalition's credibility on the market. A well-known result for normal distributions is that the cumulative distribution function can be written as :

$$Pr[P_{ij} \leq P^{CRCT}] = \frac{1}{2}\left[1 + erf\left(\frac{P^{CRCT} - \mu_{ij}}{\sigma_{ij}\sqrt{2}}\right)\right] \tag{6}$$

The amount of risks a given coalition is willing to take depends on a lot of things, among wich its capacity to compensate for under-producing (using batteries, backup generators...). Selecting the right contract value appears thus as an interesting problem on its own that we plan to investigate in future works. In order to keep the present paper in a reasonable length, we simplify the contract value selection problem a little bit by giving some responsabilities to a third party named the grid operator. The role of the grid operator is to constrain the market entry with an absolute low bound on the contract value and a upper limit relative to the coalition. In other words, the grid operator restrict the market to coalitions able to propose both sufficiently high and sufficiently credible contract values. More formally, let $\phi \in [0, 1]$ be a threshold fixed by the grid operator as a maximum value for the probability of underproducing. The highest contract value that a coalition can propose is thus $P_S^{CRCT\star}$ such that $Pr[P_{ij} \leq P_S^{CRCT\star}] = \phi$. In the gaussian example, this translates by coalition $\{i, j\}$ announcing :

$$P_S^{CRCT\star} = \mu_{ij} + \sqrt{2}\sigma_{ij}erf^{-1}(2\phi - 1) \tag{7}$$

This is the best contract value that the coalition S can afford giving the stability policy ϕ of the grid operator, and it is computable by any coalition. Figure 2 shows how

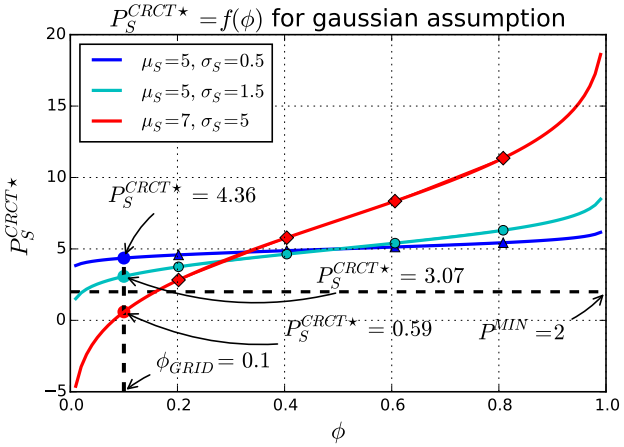

Figure 2: $P_S^{CRCT\star}$ depending on reliability parameter ϕ for gaussian distributions (see equation 7). Blue curve with triangles stands for a coalition S with an expected production of 5 units and a standard deviation of 0.5. Under a grid policy of $\phi = 0.1$, it is able to announce a contract value of $P_S^{CRCT\star} = 4.36$. The same coalition in term of expected production ($\mu = 5$), but with a higher variance ($\sigma = 1.5$, cyan curve with circles) can only afford a smaller contract value of $P_S^{CRCT\star} = 3.07$. The red curve with diamonds stands for a coalition with a higher expected production ($\mu = 7$), but with a very high unpredictability ($\sigma = 5$). For low values of ϕ, this coalition is thus heavily penalized and can only afford a contract of 0.59 units. Under grid policy ($\phi = 0.1, P^{MIN} = 2$), this last coalition is thus not allowed to enter the market (red dot below the horrizontal dashed line)
.

$P_S^{CRCT\star}$ evolves according to the reliability parameter ϕ. For illustration, the range of ϕ values is shown from 0 to 1, but in practice, only small values of ϕ really make sense : $\phi = 1$ for instance means that coalitions can announce absolutely anything since the probability of producing less than any contract value is necessarily less than one by trivial definition of a probability. As visible on figure 2, coalitions with high expected production but presenting a high unpredictability are penalized and can only afford small contracts.

In order not to overload the market with unrealistically small coalitions, the grid operator also specify a lower bound P^{MIN} on the contract values. We thus characterized a valid coalition as one satisfying the two conditions :

$$\begin{cases} Pr[P_{ij} \leq P^{CRCT}] \leq \phi \\ P^{CRCT} \geq P^{MIN} \end{cases} \tag{8}$$

On figure 2, P^{MIN} is fixed to 2 units for illustration purpose. For $\phi = 0.1$, only blue triangles and cyan circles coalitions are valid while red diamonds coalition is not.

The gaussian assumption of this small example is convenient as it allows us to write $P_S^{CRCT\star}$ analytically. Nevertheless, such assumption is rather unrealistic in practice. In the following, we keep the same framework but release this gaussian assumption. We therefore work with any observed distributions drawn from the data.

V. Coalition Formation

Now that we defined the notions of contract values and valid coalitions, we can face the problem of forming the "right" coalitions given a pool of agents. In this paper, we consider the "right" coalitions as being the ones that maximize some utility notion that, basically, indicates how much stable power can be injected in the grid. This section aims at formalizing this utility notion, while explaining how this utility will be optimized in a greedy fashion over the correlation structure of the agents.

A. Utility function

We designed our model in a way that coalitions are remunerated proportionnally to their contract values $\mathcal{U}(S) \propto P_S^{CRCT}$. That is, if λ is the unitary price rate for electricity, a coalition S injecting P_S^{CRCT} in the grid during a period $[t_0, t_k]$ earns :

$$\mathcal{U}(S) = \int_{t_0}^{t_k} \lambda P_S^{CRCT} dt = P_S^{CRCT} \lambda \Delta_t \tag{9}$$

Since λ appears, in this simple model, just a multiplicative constant, we consider for simplicity $\lambda = 1$ in the following. The utility of S is now the amount of energy S plans to inject during its contract period. Since we focused on power quantities from the begining of the present paper, and because the power injected by S in the grid is supposed to be constant over time, we simplify further the utility to its contract value : $\mathcal{U}(S) = P_S^{CRCT}$. Although concise, this formulation suffers a major drawback. It is indeed not a concave function of the coalition length, meaning that coalitions can grow as large as the number of agents allows it, without any counterparty.

Such a model, that virtually allows infinitely large coalitions, is in practice not realistic. There are indeed costs (communication costs for instance) that increase with the coalitions sizes. We take this observation into account by rescaling the utility of a coalition by its size in term of number of agents :

$$\mathcal{U}(S) = \begin{cases} \frac{1}{|S|^\alpha} \frac{P_S^{CRCT}}{P_S^{MAX}}, \ if \ S \ is \ valid, \\ 0, \ if \ S \ is \ not \ valid \end{cases} \tag{10}$$

Where α controls to what extent the size of the coalition impacts its utility, and P_S^{MAX} is a normalizing factor. P_S^{MAX} is the maximum production possible, which only occurs when all consumptions are null and all generators produce their maximum power.

Based on $\mathcal{U}$, the marginal contribution of an agent i can be expressed as $\delta_S(i) = \mathcal{U}(S + \{i\}) - \mathcal{U}(S)$. A coalition S has thus an interest in adding an additional agent i if this marginal contribution is positive :

$$\delta_S(i) \geq 0 \Leftrightarrow P_{S+\{i\}}^{CRCT} \geq P_S^{CRCT} \left(\frac{|S|+1}{|S|} \right)^\alpha \frac{P_{S+\{i\}}^{MAX}}{P_S^{MAX}} \tag{11}$$

Clearly, as α impacts the sizes of the coalitions, it should not be selected randomly. In order to have an initial guess, we use a mean field approximation in a gaussian case (as in section IV). In this section, $\bar{x}$ denote the mean approximation of quantity x. Recall that we have N agents and we wish to form N_{COAL} coalitions. We denote by $\bar{N} = \lfloor \frac{N}{N_{COAL}} \rfloor$ the expected mean number of agents in a coalition. If $\mathcal{U}(S)$ favors coalitions of size $\bar{N}$, then :

$$\left[\frac{\partial U}{\partial |S|} \right]_{|S|=\bar{N}} = 0 \tag{12}$$

Solving this equation for α in the mean field approximation for the gaussian case yields :

$$\alpha_{\bar{N}}^\star = \frac{0.7\bar{\sigma}(\bar{\rho}-1)erf^{-1}(2\phi-1)}{\bar{\mu}\sqrt{\bar{N}(\bar{\rho}\bar{N}-\bar{\rho}+1)} + 1.4\bar{\sigma}erf^{-1}(2\phi-1)(\bar{\rho}\bar{N}-\bar{\rho}+1)} \tag{13}$$

Clearly, if $\alpha = 0$, whatever the size of the coalition, additional agents are added as long as they increase the contract value. Values of α greater than zero implies that additional agents must improve the contract value by a certain coefficient that decreases and converges to one as the size of the coalition is growing. For instance, if $\alpha = 1$ and $|S| = 1$, an additional agent must first double the contract value to be included, the next agent will then have to increase it by a factor of 1.5, and so on... Obviously, α impacts directly the sizes of the coalitions formed. In the following, we consider the case where $\alpha = 1$.

As can be pointed out, the purpose of $\mathcal{U}$ is not a study of coalitions stability against player defection, wich game theory provides a lot of tools for. This is indeed a problem on its own. The redistribution of a coalition's utility in terms of individual payoffs will therefore not be considered directly in this paper. $\mathcal{U}$ can be rather interpreted as a measure of

how good a given coalition is according to our criteria since it favors small coalitions with a good production to risk ratio.

B. Representing the correlation structure

The previous section explained that coalitions with small variances in their production probability distributions are more likely to afford high contracts. Furthermore, the gaussian example highlighted that the correlation structure of the agents plays an important role in finding stable coalitions. Usually, this correlation structure is formalized with a covariance matrix or a correlation matrix that contains all the correlation coefficients between the agents : $M = (\rho_{ij})_{\forall i,j \in \mathcal{A}^2}$.

In the following, we use two opposite distance metrics :

$$\begin{cases} d^1_{ij} = 1 - \rho^2_{ij}, \\ d^2_{ij} = \rho^2_{ij} = 1 - d^1_{ij} \end{cases} \tag{14}$$

Clearly, d^1 (resp. d^2) maps two correlated series as close points (resp. distant) while two uncorrelated series are distant (resp. close). These metrics enable us to compute a correlation graph $G_1 = (\mathcal{A}, E_1)$ and a "de-correlation" graph $G_2 = (\mathcal{A}, E_2)$. For any i and j, the weight of the edge e_{ij} is d^1_{ij} in G_1 and d^2_{ij} in G_2.

Selecting the right filter ϵ is an important point since it affects the correlation structure which affects in turn the formation of the coalitions. Unfortunately, there seems to be no clear consensus in the litterature on how to select such a threshold. In our situation, we want to generate N_{COAL} coalitions from cliques in the graph. We need therefore at least N_{COAL} cliques of a given size to start. Besides, since we consider coalitions as disjoint sets, the starting cliques should be non overlapping. We thus select the optimal threshold as :

$$\epsilon^\star = min_{\epsilon \in [0,1]} \{\epsilon \ s.t. \ |\Theta_k(G^\epsilon_2)| \geq N_{COAL}\} \tag{15}$$

Where G^ϵ_2 is the de-correlation graph G_2 filtered by ϵ, and $\Theta_k(G)$ is the set of non overlapping cliques of size k in a given graph G. In other words we select ϵ as the smallest threshold posible such that the filtered de-correlation graph contains at least N_{COAL} non overlapping cliques of size k. The existence of $\epsilon^\star$ as defined in equation 15 is not guaranteed. The users has indeed to provide consistent values of N_{COAL} or k compared to the size of the agent population $\mathcal{A}$.

Correlation can be seen as cosine of angles in L^2, hence even if there is no strict transivity relation for correlation, there is, to a certain extent, some partial notion of it. More precisely, if a, b, and c are three items such that $\rho_{ab} > \delta$ and $\rho_{bc} > \delta$, then we know, by the cosine addition formula[1], that $\rho_{ac} > 2\delta^2 - 1$. That is, if a and b are strongly correlated ($\rho_{ab} > 0.9$) and b and c are also strongly correlated ($\rho_{bc} > 0.9$), then there is a high probability for a and c of being strongly correlated ($\rho_{ac} > 0.62$). This is one of the reasons why searching for clusters in correlation graphs, that is, clustering according to the correlation, makes sense.

Nevertheless de-correlation seems like a more complex concept than correlation in the sense that there is not even a partial notion of transitivity when it comes to it. As expected, the clustering coefficients of G_1 is much higher than the one of G_2. This can be seen as another formulation of Onnela's study on the structural roles of weak and strong links on correlation graphs. Strong links, accounting for strong correlation relationships, are responsible for the clustering, while weak links provide the connectivity between clusters. Searching for clusters in G_2 and hoping that this strategy will provide a nice coalition structure of internally uncorrelated coalitions seems thus pointless.

Consider now a clique in G_2, which is a complete subgraph of G_2. This is indeed a structure of interest for our purpose. Since there is a link for every pairs of nodes, we know, by construction, that a clique has a mean correlation and a maximum correlation less than ϵ. More formally, let

$$\begin{cases} \bar{\rho}_S = \frac{2}{|S|(|S|-1)} \sum_{i \in S} \sum_{j \in S, j>i} \rho_{ij}, \\ \rho^>_S = MAX_{(i,j) \in S^2}(\rho_{ij}) \end{cases} \tag{16}$$

be the mean correlation of coalition S and the maximum correlation value in S. If S is a clique in G_2 and $\epsilon \in [0,1]$ is the threshold used for filtering G_2, then $\bar{\rho}_S \leq \frac{2}{|S|(|S|-1)} \sum_{i \in S} \sum_{j \in S, j>i} \epsilon$, that is, $\bar{\rho}_S \leq \epsilon$, and $\rho^>_S \leq \epsilon$.

Because of this de-correlation property, cliques in G_2 appear as good candidates for coalitions. Nevertheless, by doing so, coalitions are often small and only formed based on the underlying correlation structure of the agents, and not on power generation capacity. It is possible that adding agents to these coalitions has the combined effect of increasing the production while decreasing its stability. The question revolves around measuring the benefits of this production surplus compared to the disadvantage of a higher unstability. Hopefully, this can be quantified exactly with the marginal benefit developped in the previous section.

C. Coalition formation algorithm

The algorithm takes inputs from :

- **The agents :** historical series of available productions P_i,
- **The grid operator :** market entrance policy (P^{MIN}, ϕ),
- **The "user" :** Number of desired coalitions N_{COAL} and size of starting cliques k.

The first steps consists in computing the de-correlation graph G_2 as well as the optimal threshold $\epsilon^\star$. Cliques of size k in $G^{\epsilon^\star}_2$ are considered as coalition seeds. The next step is a local greedy improvement over the correlation structure represented by $G^{\epsilon^\star}_2$. Cliques add alternatively the node $i^\star$ in their neighborhood that yields the best marginal benefit $MAX_{i \in N(clique)} \delta_{clique}(i)$ where $N(clique)$ is the neighborhood of a given clique. This addition occurs only if $i^\star$ is not already involved in another coalition, and if $\delta_{clique}(i^\star) \geq 0$, meaning that utilities are increasing. The algorithm stops when all nodes are distributed in a coalition or when the global utility stops increasing. See the details in algorithm.

[1] $cos(a+b) = cos(a)cos(b) - sin(a)sin(b)$

Data: P_i series,
Grid policy (P^{MIN}, ϕ),
Desired number of coalitions N_{COAL},
size of starting cliques k
Result: $CS = \{S_1, ..., S_{N_{COAL}}\}$
Compute $G_2^{\epsilon^\star}$;
Find the N_{COAL} cliques in $G_2^{\epsilon^\star}$;
while $\mathcal{U}(CS)$ *is improving* **do**
  for *each clique* **do**
    Find $i^\star$;
    if $\delta_{clique}(i^\star) >= 0$ **then**
      $clique \leftarrow clique \cup \{i^\star\}$;
    end
  end
end

Algorithm 1: Percolation algorithm

Data: Agent set $\mathcal{A}$,
Desired number of coalitions N_{COAL},
Maximum number of iterations $maxiter$
Result: $CS = \{S_1, ..., S_{N_{COAL}}\}$
$curriter \leftarrow 0$;
$CS^\star \leftarrow \emptyset$;
while $curriter < maxiter$ **do**
  $CS \leftarrow SelectRandomCS()$;
  if $\mathcal{U}(CS) > \mathcal{U}(CS^\star)$ **then**
    $CS^\star \leftarrow CS$;
  end
  $curriter \leftarrow curriter + 1$;
end
return $CS^\star$

Algorithm 2: Random algorithm

Data: P_i series,
Desired number of coalitions N_{COAL},
search step size $\beta << 1$
Result: $CS = \{S_1, ..., S_{N_{COAL}}\}$
$\epsilon \leftarrow 1$;
$CS \leftarrow \emptyset$;
while $|CS| < N_{COAL}$ **do**
  Compute G_1^ϵ ;
  $CS \leftarrow computeClusters(G_1^\epsilon)$;
  if $|CS| = N_{COAL}$ **then**
    return CS;
  end
  else
    $\epsilon \leftarrow \epsilon - \beta$;
  end
end

Algorithm 3: Correlated algorithm

VI. Results

The algorithm presented in the previous section is supposed to generate a given number of coalitions that have good utilities, and therefore, high contract values for relatively small sizes. Nevertheless, as it comprises mainly of a greedy optimization based on local improvements, the probability that the algorithm finds the optimal coalitions set is very low. Actually, it is not obvious that a strict optimum exists at all. Besides, there is no, to our knowledge, state of the art algorithm that aggregate uncorrelated series in an optimum way.

In order to have an idea about the algorithm's quality, we compare its results with :

- **Random :** This algorithm is basically a random search over the coalition structures space. It analyses a given number of structures and returns the best that it has encountered so far. See algorithm.
- **Correlated :** This is the complete opposite of our algorithm. It basically uses the correlation graph G_1 and performs community detection on it. The resulting coalitions have thus very high internal correlations. We thus expect this algorithm to perform very bad compared to the others. See algorithm.
- **K-means :** This is simply a K-means clustering of the agents P_i's series. The goal of presenting this algorithm is to show that K-means do not consider correlations explicitely and is therefore not well suited for our purpose.

The coalitions formed according to these algorithms will be compared using two criteria : the normalized global utility of the coalitions formed $\bar{\mathcal{U}}(CS) \in [0, 1]$, and the percentage of valid coalitions that satisfy the grid conditions $\tau(CS) \in [0, 1]$:

$$\bar{\mathcal{U}}(CS) = \frac{1}{N_{COAL}} \sum_{S \in CS} \mathcal{U}(S) \tag{17}$$

$$\tau(CS) = \frac{|\{S \in CS,\ s.t\ \mathcal{U}(S) > 0\}|}{N_{COAL}} \tag{18}$$

VII. Conclusion