Top Tagging: A Method for Identifying Boosted Hadronically Decaying Top Quarks

David E. Kaplan, Keith Rehermann, Matthew D. Schwartz, and Brock Tweedie

Department of Physics and Astronomy, Johns Hopkins University, Baltimore, Maryland 21218, USA
(Received 9 June 2008; published 2 October 2008)

A method is introduced for distinguishing top jets (boosted, hadronically decaying top quarks) from light-quark and gluon jets using jet substructure. The procedure involves parsing the jet cluster to resolve its subjets and then imposing kinematic constraints. With this method, light-quark or gluon jets with $p_T \simeq$ 1 TeV can be rejected with an efficiency of around 99% while retaining up to 40% of top jets. This reduces the dijet background to heavy $t\bar{t}$ resonances by a factor of $\sim 10\,000$, thereby allowing resonance searches in $t\bar{t}$ to be extended into the all-hadronic channel. In addition, top tagging can be used in $t\bar{t}$ events when one of the top quarks decays semileptonically, in events with missing energy, and in studies of b-tagging efficiency at high p_T.

The Large Hadron Collider (LHC) is a top-quark factory. The millions of top quarks it produces will provide profound insights into the standard model and its possible extensions. Most of the top quarks will be produced near threshold and can be identified using the same kinds of techniques applied at the Tevatron—looking for the presence of a bottom quark through b tagging, identifying the W boson, or finding three jets whose invariant mass is near m_t. However, some of the top quarks produced at the LHC will be highly boosted. In particular, almost every new physics scenario that addresses the hierarchy problem will include new heavy particles which decay to top quarks (such as KK gluons in Randall-Sundrum models, squarks in supersymmetry, top primes in little Higgs models, etc.). If their masses are even a factor of a few above the top-quark mass, the top quarks that they produce will decay to collimated collections of particles that look like single jets. In this case, the standard top-quark identification techniques may falter: b tagging is difficult because the tracks are crowded and unresolvable, the W decay products are not always isolated from each other or from the b jet, and the top jet mass may differ from m_t due to an increased amount of QCD radiation.

In most studies of $t\bar{t}$ resonances, emphasis is placed on the channel in which one top quark decays semileptonically (to an electron or muon, a neutrino, and a b jet) and the other hadronically [1,2]. This avoids having to confront the large dijet background to all-hadronic $t\bar{t}$. However, these studies need to assume that the lepton can be isolated, which often excludes the electron channel, and that at least one b jet is tagged, which is difficult at high p_T [3]. The hard muon tag alone already discards 90% of the $t\bar{t}$ events. So one would like to be able to use the all-hadronic channel without b tags. In this Letter, we introduce a practical and efficient method for tagging boosted hadronically decaying top quarks.

A top quark's dominant decay mode is to a b quark and a W boson with the W subsequently decaying to two light quarks. The three quarks normally appear as jets in the calorimeter, but for highly boosted top quarks these jets may lie close together and may not always be independently resolved. For example, a zoomed-in lego plot of a typical top jet is shown in Fig. 1. It displays energy deposited in the calorimeter versus pseudorapidity η and azimuthal angle ϕ. The three quark jets show up clearly by eye, but it is easy to see how the number of jets identified by conventional clustering would be highly variable and strongly dependent on the jet-resolution parameter.

The natural direction for finding boosted top quarks is to look into subjet analysis and other measures of the energy distribution in the events. A recent ATLAS report [4] explored the possibility by cutting on the jet mass and the y_{cut} variables associated with the k_T algorithm. This method was not found to be strong enough to filter $t\bar{t}$ events from the enormous dijet background (see also Ref. [5] for another approach with similar efficiencies).

The key to efficient top tagging is in isolating features of QCD which control the background from features particular to the top quark. As can be seen in Fig. 1, boosted top-quark events look like single jets with three resolvable
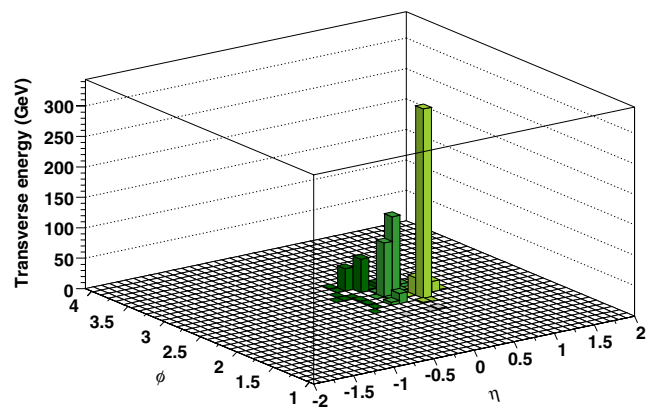

FIG. 1 (color online).　A typical top jet with a p_T of 800 GeV at the LHC. The three subjets after top tagging are shaded separately.

subjets in a small region of the calorimeter. These subjets are separated by angular scales of order $\sim 2m_t/p_T$ and so remain distinguishable from one another up to p_T's of roughly 2 TeV for a calorimeter cell size of 0.1. In QCD, on the other hand, a typical high-p_T jet starts as a single hard parton, which subsequently cascades into a high multiplicity of soft and collinear particles. Most of these particles cannot be resolved by the real calorimeter, as they tend to fall into a single cell or a set of adjacent cells. In order to look like a decayed top quark, a hard parton must at least undergo two branchings at somewhat large angles and energy sharings, which is relatively rare, as we will see. The primary task, then, is to isolate events with three hard, nearby subjets. Subsequently, we may exploit the full 3-body kinematics of top decay to construct additional discriminating variables.

In order to avoid the pitfalls mentioned above for fixed-size jet clustering, we first cluster an event using a large jet radius to capture all of the potential substructure and then iteratively decluster each jet to search for subjets. Similar ideas have been employed by Butterworth *et al.* to extract substructure in Higgs jets [6] and W jets [7,8], and part of our algorithm is an adaptation of their method.

The top tagging algorithm is as follows.

First, particles are clustered into jets of size R. For this step, we use the Cambridge-Aachen (CA) algorithm [9,10]. This iterative procedure begins with all four-vectors in an event, as defined by the energy deposits in the calorimeter. It then finds the pair which is closest in $\Delta R = \sqrt{\Delta \eta^2 + \Delta \phi^2}$, merges it into a single four-vector, and then repeats. The procedure ends when no two four-vectors have $\Delta R < R$.

Next, each jet in the event (for $t\bar{t}$ this would be one of the hardest two) is declustered, to look for subjets. This is done by reversing each step in the CA clustering, iteratively separating each jet into two objects. The softer of the two objects is thrown out if its p_T divided by the full jet p_T is less than a parameter δ_p, and the declustering continues on the harder object.

The declustering step is repeated until one of four things happens: (1) Both objects are harder than δ_p; (2) both objects are softer than δ_p; (3) the two objects are too close, $|\Delta \eta| + |\Delta \phi| < \delta_r$, where δ_r is an additional parameter; or (4) there is only one calorimeter cell left. In case (1), the two hard objects are considered subjets. In cases (2), (3), and (4), the original jet is considered irreducible.

If an original jet declusters into two subjets, the previous step is repeated on those subjets (with δ_p still defined with respect to the original jet's p_T) resulting in 2, 3, or 4 subjets of the original jet. The cases with 3 or 4 subjets are kept, the 4th representing an additional soft gluon emission, while the 2 subjet case is rejected.

With these 3 or 4 subjets in hand, additional kinematic cuts are imposed: The total invariant mass should be near m_t, two subjets should reconstruct m_W, and the W helicity angle should be consistent with a top decay, as described below.

For our particular implementation, we simulate dijet events and $t\bar{t}$ events in the standard model at the LHC using PYTHIA V.6.415 [11]. In order to simulate the resolution of the ATLAS or CMS calorimeters, particles in each event are combined into square bins of size $\Delta \eta = \Delta \phi = 0.1$, which are interpreted as massless four-vector "particles" and inputted into the clustering routine. For jet clustering, we employ the CA algorithm as implemented in FASTJET V.2.3.1 [12]. Because more highly boosted top quarks will be more collimated, we correlate the jet clustering parameter R, the event's scalar E_T, and the two clustering parameters δ_p and δ_r as follows: For $E_T >$ 1000, 1600, and 2600 GeV, we take $R = 0.8$, 0.6, and 0.4, $\delta_p = 0.10$, 0.05, and 0.05, and $\delta_r = 0.19$, 0.19, and 0.19, respectively. Then we demand that the jets be hard by putting a cut on the jet p_T scaled by the event's scalar E_T: $p_T > 0.7\frac{E_T}{2}$. Both jets must also satisfy the absolute constraints $p_T > 500$ GeV and $|\eta| < 2.5$ to be considered for analysis.

Next, we perform the subjet decomposition, demanding 3 or 4 subjets, as described above. For jets with $p_T < 1000$ GeV, we then ask that the invariant mass of the sum of the subjet four-vectors be within 30 GeV of the top mass (145–205 GeV) and that there exist two subjets which reconstruct the W mass to within 15 GeV (65–95 GeV). Harder jets will have broader mass distributions, due to increased radiation from QCD. Thus, if a jet has $p_T > 1000$ GeV, we shift the upper ranges of top and W mass cuts to $p_T/20 + 155$ GeV and $p_T/40 + 70$ GeV, respectively. Finally, we demand that the W helicity angle satisfy $\cos\theta_h < 0.7$, as we now explain.

The helicity angle is a standard observable in top decays, used to determine the Lorentz structure of the top-W coupling [13]. It is defined as the angle, measured in the rest frame of the reconstructed W, between the reconstructed top quark's flight direction and one of the W decay products. Normally, it is studied in semileptonic top decays, where the charge of the lepton uniquely identifies these decay products. In hadronic top decays there is an ambiguity which we resolve by choosing the lower p_T subjet, as measured in the lab frame. (Other choices are possible and make little difference on the final efficiencies.) For top jets, the distribution is basically flat: Since the W decays on-shell, its decay products are almost isotropically distributed in the W rest frame. In contrast, for light-quark or gluon jets, the distribution diverges (at the parton level) as $1/(1 - \cos\theta_h)$. This corresponds to a soft singularity in the QCD matrix elements for emitting an additional parton. Example distributions are shown in Fig. 2. The qualitative features we understand analytically at the parton level are clearly visible after showering and hadronization. Other observables sensitive to the soft singularity are possible [5] and will give similar signal/background enhancements.
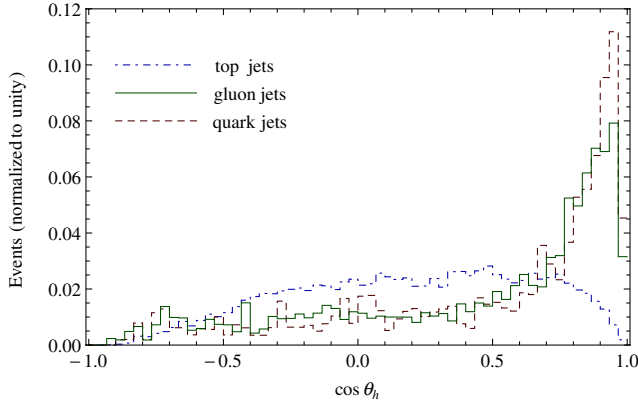
FIG. 2 (color online).   Distribution of helicity angle for top jets, gluon jets, and light-quark jets for $p_T > 700$ GeV. These distributions are after the subjet requirement, top-quark mass cut, and W mass cut have been imposed.

To check the efficacy of this method, we calculate the efficiency for correctly tagging a top jet ϵ_t and the efficiencies for mistagging light-quark or gluon jets as top quarks ϵ_q and ϵ_g, respectively. These are shown in Fig. 3. There are a few important qualitative observations one can make about this plot. For very large p_T, the top tagging efficiency goes down. This is because these jets are so highly boosted that the calorimeter can no longer distinguish the subjets. As p_T goes below 900 GeV, the top tagging efficiency also decreases. This is due to some of the top jets becoming too fat for the initial $R = 0.8$ clustering. (This somewhat tight choice was made to suppress the mistag efficiency, which grows faster than the top tag efficiency with increasing R.) Examples of the sequential effects of the individual cuts are shown in Table I. The clustering R's and kinematic cuts can be varied to increase the tagging and mistagging efficiencies, as desired for a particular $S/\sqrt{B}$ goal.
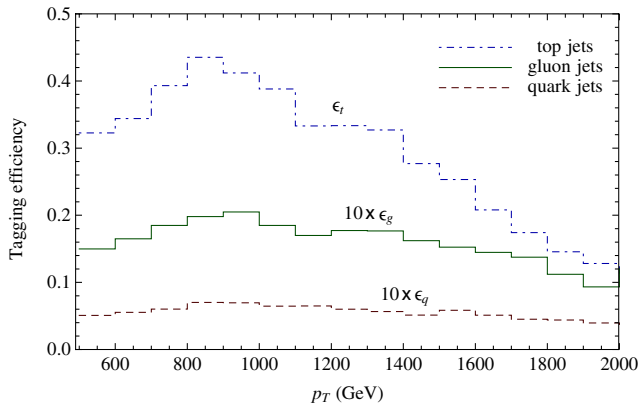

FIG. 3 (color online).   The efficiencies for correctly tagging a top jet (ϵ_t) and mistagging a gluon jet (ϵ_g) or a light-quark jet (ϵ_q). The quark and gluon efficiencies are of order 1% and have been scaled in the plot by a factor of 10 for clarity.

One important concern is whether the Monte Carlo simulation generates the $t\bar{t}$ and dijet distributions correctly. To test this possibility, we redid our analysis using samples generated with various shower parameters, with the "new" p_T-ordered dipole shower in PYTHIA and with HERWIG V.6.510 [14]. We find a 50% variation in ϵ_q and ϵ_g and a negligible change in ϵ_t. We also ran PYTHIA with multiple interactions and initial state radiation turned off, individually and together. Effects on ϵ_q and ϵ_g are at the 10% level or less, indicating that the QCD jet substructure relevant for top tagging is mostly controlled by final state parton branchings.

One might also be worried about whether, since we are looking at multi(sub)jet backgrounds, it would be important to include full matrix element calculations. However, since the events are essentially two jet events, the substructure is due almost entirely to collinear radiation, which the parton shower should correctly reproduce [15]. To confirm this, we have also simulated background events using MADGRAPH V.4.2.4 [16]. Using events with $2 \rightarrow 4$ matrix elements in a region of phase space where 1 parton recoils against 3 relatively collinear partons, we repeated our analysis without showering or hadronization. The resulting mistag efficiencies were consistent with those from the PYTHIA study to within 10%, which provides justification for both the parton shower approximation and the robustness of our algorithm.

One possible way to verify the Monte Carlo predictions for jet substructure would be to use data directly. For example, the efficiency of the top tagging algorithm can be calibrated by comparing the rate for $t\bar{t}$ events where one top quark decays semileptonically with the rate in the all-hadronic channel. The background rejection efficiency can also be studied by looking in sidebands where the jet invariant mass is not close to m_t.

Top tagging may be particularly useful in the search for new physics in $t\bar{t}$ resonances. In the all-hadronic channel, the biggest background for $t\bar{t}$ is dijets, so in Fig. 4 we show the dijet and $t\bar{t}$ invariant mass distributions before and after top tagging both jets. It is evident that, after top tagging, the dijet sample is reduced to the level of the $t\bar{t}$ sample. As

TABLE I.   Incremental efficiencies for top, gluon, and light-quark jets passing the subjets, invariant mass, and helicity angle cuts for jets in three different p_T windows.

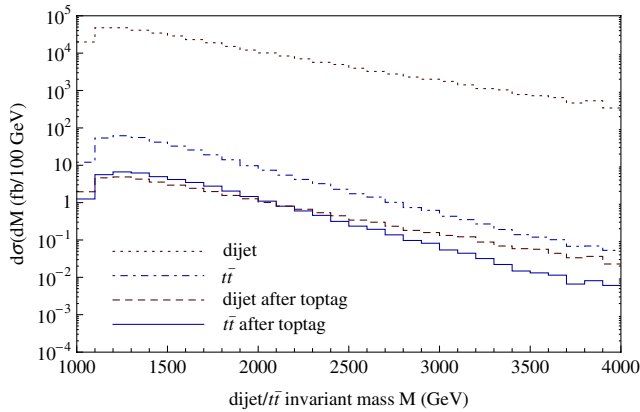|	p_T (GeV)	Subjets	m_t	m_W	θ_h
ϵ_t	500–600	0.56	0.43	0.38	0.32
	1000–1100	0.66	0.52	0.44	0.39
	1500–1600	0.40	0.33	0.28	0.25
ϵ_g	500–600	0.135	0.045	0.027	0.015
	1000–1100	0.146	0.054	0.032	0.018
	1500–1600	0.083	0.038	0.025	0.015
ϵ_q	500–600	0.053	0.018	0.011	0.005
	1000–1100	0.063	0.023	0.013	0.006
	1500–1600	0.032	0.015	0.010	0.006

FIG. 4 (color online).   Effect of the top jet tag on standard-model $t\bar{t}$ and dijet distributions at the LHC. Both the t and the $\bar{t}$ decay hadronically, and no b tagging is used. With top tagging, a strongly produced $t\bar{t}$ resonance (not shown) would stand out clearly over the background in this channel.

an example application, in certain Randall-Sundrum models [17,18] KK gluons decay dominantly to $t\bar{t}$. It has been shown that, if one can isolate the $t\bar{t}$ events, the resonance will stand out as a clean peak over the standard-model $t\bar{t}$ background [1,2,19]. Since top tagging can reduce the dijet background to the size of the $t\bar{t}$ background, $t\bar{t}$ resonance searches can be done in the all-hadronic channel for resonances up to a few TeV.

There are many applications for top tagging besides $t\bar{t}$ resonances searches. For example, a common new physics signal is $t\bar{t}$ pairs in association with missing energy [20]. This may happen, for instance, in supersymmetry when heavy top squark pairs decay to highly boosted top quarks and neutralinos. Top tagging can not only reduce the standard-model backgrounds in this context, but it can also help distinguish top jets from light-quark jets in any signal event, which may be helpful in studying the flavor structure of the new physics. In addition, top tagging could potentially be applied in searches for single top-quark events where exactly one top jet is required. Finally, our technique could be used as a handle for measuring b-tagging efficiency at high p_T.

In conclusion, we have demonstrated that it is possible to distinguish highly energetic top quarks from standard-model backgrounds at the LHC. With efficiencies $\epsilon_t \sim 40\%$ and $\epsilon_q \sim \epsilon_g \sim 1\%$, top tagging is better than

b-tagging at high p_T. Top jets can now be considered standard objects for event analysis at the LHC, as b jets are at the Tevatron.

[1]   K. Agashe, A. Belyaev, T. Krupovnickas, G. Perez, and J. Virzi, Phys. Rev. D **77**, 015003 (2008).
[2]   U. Baur and L. H. Orr, Phys. Rev. D **77**, 114001 (2008).
[3]   See, for example, L. March, E. Ros, and B. Salvachúa, Report No. ATLAS-PHYS-PUB-2006-002.
[4]   G. Brooijmans, Report No. ATL-COM-PHYS-2008-001, 2008.
[5]   J. Thaler and L. T. Wang, J. High Energy Phys. 07 (2008) 092.
[6]   J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, Phys. Rev. Lett. **100**, 242001 (2008).
[7]   J. M. Butterworth, B. E. Cox, and J. R. Forshaw, Phys. Rev. D **65**, 096014 (2002).
[8]   J. M. Butterworth, J. R. Ellis, and A. R. Raklev, J. High Energy Phys. 05 (2007) 033.
[9]   Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, J. High Energy Phys. 08 (1997) 001.
[10] M. Wobisch and T. Wengler, arXiv:hep-ph/9907280.
[11] T. Sjostrand, S. Mrenna, and P. Skands, J. High Energy Phys. 05 (2006) 026.
[12] M. Cacciari and G. P. Salam, Phys. Lett. B **641**, 57 (2006).
[13] T. Chwalek *et al.* (CDF Collaboration and D0 Collaboration), arXiv:0705.2966.
[14] G. Corcella *et al.*, J. High Energy Phys. 01 (2001) 010.
[15] See, for example, E. Norrbin and T. Sjostrand, Nucl. Phys. **B603**, 297 (2001).
[16] J. Alwall *et al.*, J. High Energy Phys. 09 (2007) 028.
[17] L. Randall and R. Sundrum, Phys. Rev. Lett. **83**, 3370 (1999).
[18] See, for example, K. Agashe, R. Contino, and A. Pomarol, Nucl. Phys. **B719**, 165 (2005).
[19] B. Lillie, L. Randall, and L. T. Wang, J. High Energy Phys. 09 (2007) 074.
[20] T. Han, R. Mahbubani, D. G. E. Walker, and L. T. E. Wang, arXiv:0803.3820.